UNIVERSITÄT ZU LÜBECK

From the Institute of Information Systems
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Ralf Möller

Semantic Assets: Latent Structures for Knowledge Management

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences

Submitted by

Sylvia Melzer
from Hamburg
Lübeck, 2018

First referee: Prof. Dr. Ralf Möller

Second referee: Prof. Dr. Christian Willi Scheiner

Date of oral examination: June 21, 2018

Approved for printing. Lübeck, August 15, 2018

Abstract

In standard information retrieval systems, queries can be specified with different languages (string patterns, logical formulas, and so on). It is well known that it is hard to simultaneously maximize quality measures for query answers, such as, e.g., precision and recall.

Retrieval of documents with high recall, while maintaining at least a decent precision level, is a frequent problem in knowledge management (KM) contexts based on information retrieval (IR) processes, e.g., for information association in business contexts. Similar to making implicit knowledge explicit, deriving explicit symbolic content descriptions is important in KM tasks.

In this thesis we show how explicit symbolic descriptions can be combined with implicit holistic content representations known from information retrieval in order to support knowledge management processes in general, and information association based on IR query answering in particular. The methodology exhibited in this thesis is verified using representative examples, and validated with feasibility and effectiveness studies.

Zusammenfassung

In Standard-Information-Retrieval-Systemen können Anfragen mit verschiedenen Sprachen (Strings, logische Formeln usw.) gestellt werden. Es ist bekannt, dass es schwierig ist, verschiedene Qualitätsmaße gleichzeitig für die Suche nach Antworten zu maximieren, wie beispielsweise Trefferquote und Präzision.

Das Retrieval von Dokumenten mit einer hohen Trefferquote, während zumindest eine angemessene Präzision erhalten bleibt, ist ein häufiges Problem in Wissensmanagement-Kontexten basierend auf Prozessen des Information-Retrieval (IR), z.B. zur Zusammenführung von Informationen in einem Geschäftskontext. Ähnlich wie implizites Wissen explizit gemacht wird, ist die Ableitung von expliziten symbolischen Inhaltsbeschreibungen ein wichtiger Aspekt im Wissensmanagement.

In dieser Arbeit wird untersucht, wie explizite symbolische Beschreibungen mit impliziten ganzheitlichen Inhaltsrepräsentationen beim IR kombiniert werden können, um Prozesse im Allgemeinen zu unterstützen und besonders die Vereinigung von Informationen basierend auf der IR-Anfragebeantwortung. Die in dieser Arbeit vorgestellte Methodik wird unter Verwendung von repräsentativen Beispielen verifiziert; und die Durchführbarkeit und Qualität durch Studien validiert.

Contents

Chapter 1

Introduction

Modern industrial nations are based on the knowledge of employees working in companies, and thus, knowledge management (KM) has become a major concern for achieving productivity advantages. Computer systems are used to support KM in various application contexts. However, no clear definition of KM has yet emerged. It is a central idea of this thesis to substantiate notions of knowledge management based on information retrieval (IR).

Typically, knowledge management processes are based on content management (CM) systems. In CM systems, content is stored, organized, and supplemented with metadata. Among simple data for authors, characters, publishers, and so on, nowadays metadata contain feature-based as well as *symbolic content descriptions* (also called *symbolic representations*, see Figure 1.1), which, for instance, can be represented via logic-based techniques [Kay11, EP11]. Applications exploit symbolic content descriptions in various ways. For example, in the semantic web, content descriptions are used for finding documents, images, videos, or persons. Search requests are specified by posing *queries* in query languages based on string patterns, logical formulas, and so on.

For matching queries with content as well as with content descriptions, each query language has its pros and cons [Mel06]. For most purposes, string patterns have a high recall but do not lead to high precision.[1] In practice it is difficult to maximize precision and recall simultaneously. Until now, large-scale

[1] Recall is defined as the number of relevant items retrieved divided by the number of relevant items in the repository, while precision is defined as the number of relevant items retrieved divided by the overall number of retrieved items.

information retrieval processes are rarely based on symbolic content descriptions for matching queries with content [VH05]. Google's Knowledge Vault (KV) uses symbolic descriptions in order to help the user create useful follow-up queries [DMG+14]. It is also possible that so called *holistic content descriptions* (e.g., TF.IDF matrices) and corresponding similarity measures are used for query answering [SWY75, MRS08a]. Matches on holistic content descriptions can be realized efficiently, for example, by utilizing nearest-neighbor algorithms.

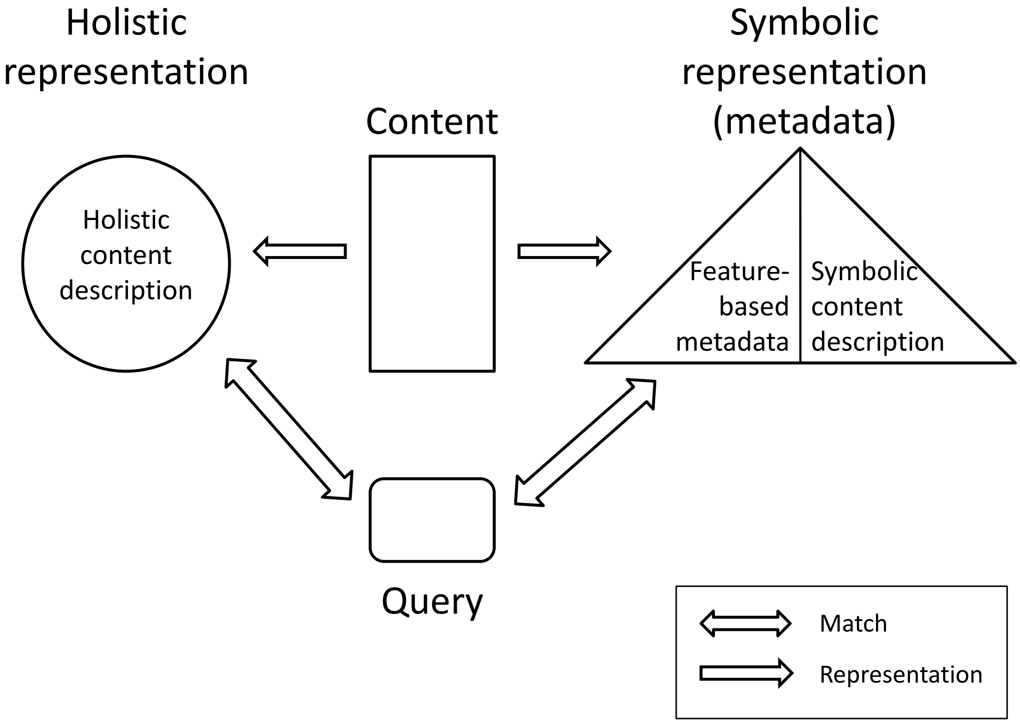

Holistic representation
Content
Symbolic representation (metadata)
Holistic content description
Feature-based metadata
Symbolic content description
Query
Match
Representation

Figure 1.1: **Content representation formats.** Description see text.

In short and slightly exaggerating, holistic representations lead to high recall and low precision, and symbolic representations tend to be characterized by low recall and high precision [BMM92, PKM+07a, PKM+07b]. In the symbolic context, it is desirable to increase recall while at least maintaining precision. In the holistic context the goal is to increase precision. Our hypothesis is that this kind of improvement could be achieved by systematically combining symbolic and holistic content descriptions. In the literature there are ideas to combine symbolic and holistic content descriptions [WC16], but to the best of our knowledge, a combination of both kinds of content descriptions has been investigated as an extension to the standard boolean model [Sal83, SM86]. This early work, however, is based on feature-based metadata only (e.g., resolution of images, video encoding, etc.) and not on symbolic content descriptions (see Figure 1.1). Many contributions are based on a holistic approach such as latent semantic indexing (LSI) presented in [DDF$^+$90a], syntax- or dependency-based models presented in [PL07a], inducing latent semantic relations for structured distributional semantics presented in [JH14], or the distributed holistic clustering approach for linking many data sources in order to enable an effective and efficient clustering of entity sets from many data sources presented in [NGMR17]. Holistic methods for quantifying and categorizing semantic similarities are also called distributional semantics. Distributional semantic models vary w.r.t. the usage of frequency weighting, dimension reduction, similarity measure, and differ in data representations [RG65, Lin98, PL07b, RM10, RB12, LG14, FPBP16, SKI16, PRF$^+$17].

It is a central idea of this thesis to suggest ways for systematically combining symbolic and holistic content descriptions in order to increase recall while at least maintaining precision.

In the following, retrieved documents which are relevant will be called *high-quality documents*. Retrieval of high-quality documents is a frequent task in KM contexts, in the sense that the documents themselves or, in some applications, their authors are subject to further steps in KM processes. However, we consider a use case in which finding documents might be a problem in case that there is no direct match with simple queries. Consequently, queries need to reformulated, which usually is a rather difficult task for users. This is true for pattern-based as well as logic-based queries [Mel06, SI09]. Indeed, if there are at least some query results due to a symbolic search, we argue that these results can be analyzed and exploited for detecting relevant additional material in order to find high-quality documents based on holistic search. While an increase of recall might indeed be the result, the decrease in precision needs to be controlled. It is our hypothesis that this idea can be realized with a systematic comparison of the symbolic descriptions of the initial results obtained based on a holistic retrieval (see Figure 1.1).

1.1 Research Objectives

The main objectives of this thesis are the substantiation of knowledge management notions and the enhancement of recall for queries in a KM context, while at least maintaining precision, by suggesting ways for the combination of holistic and symbolic content descriptions.

For deriving symbolic descriptions we further investigate fusion of multimodal representations, and we extend logic-based interpretation of content. Logic-based techniques such as the non-standard inference service *A-box abduction* have been studied and developed for representing symbolic content descriptions [Kay11, EP11]. Symbolic representations (annotations) describe documents, images, videos, or persons and can be seen as an *interpretation* of content. The combination of several interpretation results is called *fusion*. First investigations on a fusion process for multi-modal interpretation results were done in [Kay11] in order to fuse symbolic content descriptions of different parts of a multimedia document such that precision of retrieval results is increased. In this process, the annotations of a multimedia document will be conjoined and the individuals of one part will be identified with others if the individuals describe the same real-world entity. However, the fusion algorithm in [Kay11] fuses only individuals from different modalities, and due to simple forward-chaining rules only in very specific situations. Thus, a further objective of this work is to expand the view on fusion such that no situation specific rules need to be specified. For this purpose the so-called *A-box difference operator* [MGK$^+$14] is employed in order to define a new fusion algorithm.

The retrieval of documents using holistic content descriptions is well established [MRS08a], and the investigation in this thesis will be based on *latent semantic indexing* (LSI), such that we have a technology for holistic search with high recall [Gee03a]. The challenge is to combine LSI with symbolic retrieval. A combination of approaches for representing content symbolically and holistically poses exceptional technological challenges in order to increase precision on the one hand, and recall on the other. We argue that the combination of both approaches in a systematic way achieves better matching results. In other words, our hypothesis is that IR processes, which are based on systematically combined holistic and symbolic content descriptions for matching queries with content, can result in high recall without an associated decrease of precision. In this context it should be noted that this thesis has not the objective to present absolute numbers or performance measurements for a concrete KM application. However, the objective is to reify the relationship between symbolic and holistic representations by defining so-called *semantic assets* as a basis for building KM sytems in the future.

We argue that semantic assets are essential for knowledge management in companies. In most companies a lot of knowledge of documents and the knowledge in the people's head are available, but employees often do not know how to utilize all available knowledge. In [NKT98, Non08] it is shown how employees in a company can exploit their knowledge for innovation. Nonaka et al. define a so-called knowledge-creation process, which provides general foundations for creating knowledge. In the literature it has been argued [ES97, SMD12] that nevertheless company managers still have to define a concrete knowledge-creation process in order to develop innovative products. In other words, to apply Nonaka's ideas in a fruitful way, there is still a strong need for a formalization of KM notions. In this thesis, another research objective is to achieve a further substantiation of Nonaka's idea of knowledge-creation processes based on semantic assets.

1.2 Research Methodology

We define a methodology as a collection of related processes, methods, and tools [Est08]:

- A process is a logical sequence of tasks performed to achieve a particular objective. A process defines "WHAT" is to be done, without specifying "HOW" each task is performed.

- A method consists of techniques for performing a task, in other words, it defines the "HOW" of each task.

- A tool is an instrument that, when applied to a particular method, can enhance the efficiency of the task; provided it is applied properly and by somebody with proper skills and training. The purpose of a tool should be to facilitate the accomplishment of the “HOWs”. In a broader sense, a tool enhances the “WHAT” and the “HOW”.

This thesis pursues different research approaches w.r.t. processes, methods, and tools. An analysis of KM and CM is conducted with the focus on KM processes based on combinations of symbolic and holistic content descriptions in order to define the “WHAT”. The use and benefit of systematically combined symbolic and holistic content descriptions is investigated to improve information retrieval in a knowledge management environment. In order to achieve an improvement, this thesis presents “HOW” to combine symbolic and holistic representations profitably. Moreover, we argue that existing annotation concepts such as RDFa (see [W3C14a]) are useful, but have disadvantages w.r.t. generality, e.g., for multimodal documents. The central idea of this thesis deals with fundamental principles of combining symbolic and holistic representations in a KM application area.

The purpose of reification of the combination by defining *semantic assets* in a KM application area is the definition of KM notions using semantic assets in order to manage semantic assets in a document structure.

Schmidt and Sehring define content descriptions as *assets* by considering pairs of media content and conceptual abstractions [SS04, SS03]. Bossung and Schmidt develop a structurally rich way in order to represent and to handle multimedia content [Bos08]. Accordingly, our idea is to extend the approaches of Schmidt and Sehring, and Bossung and Schmidt in order to treat the central problem of handling knowledge with associated inferences using existing tools. Software for content management has been developed in [SS03, SS04, Seh04, Sun06, CFO10], and we argue that the approaches, for example, from Sehring presented in [Seh04] are applicable to knowledge management. Thus, the tool selection in this thesis is clarified as well as the methods and processes.

The methodology used in this thesis, with the aim profitably combining holistic and symbolic representations for information retrieval, is evaluated with representative examples from a large multimedia repository from the EU project BOEMIE.[2] In order to complement these data we also use Wikipedia pages and commercial annotation services. For presentation purposes we also

[2] Bootstrapping Ontology Evolution using Multimedia Information Extraction

use a simple handcrafted example. The evaluation results show performance of information retrieval and quality of retrieval results.

1.3 Contributions

The major contributions of this thesis are summarized as follows:

- The knowledge-creation process defined in [NKT98] is supported with formal operators.
- Research approaches are investigated in order to find advantageous ways for the combination of holistic and symbolic content representations with the result to increase recall and at least maintaining precision for information retrieval tasks in KM.
- An analysis of KM requirements is done in order to reify fundamental aspects of knowledge management systems and the management of knowledge.
- Semantic assets based on description logics are defined for the purpose of representing knowledge described in documents.
- The investigations of Sehring [Seh04] and Bossung [Bos08] are extended for defining semantic assets in order to develop a structurally rich way to represent and to handle multimedia content in a KM environment.
- A new logic-based fusion algorithm approach is presented which is used for the combination of symbolic and holistic content descriptions. The proposed logic-based fusion algorithm is based on the so-called *A-box difference operator*. The A-box difference operator is used, on the one hand, to obtain the difference of two content descriptions, and on the other hand to identify individuals which describe the same real-world entity. For the new logic-based fusion process the identification of identical individuals is required in order to fuse equal individuals with the aim to increase precision.

Some parts of this work were published previously. The following paper present use cases for computing differences of variants and show how to apply the

A-box difference operator for symbolic content descriptions for constructing knowledge representations.

- D. Arndt, S. Melzer, R. God, and M. Sieber. **Konzept zur Verhaltensmodellierung mit der Systems Modeling Language (SysML) zur Simulation varianten Systemverhaltens.** *Tagungsband zum Tag des Systems Engineering (Eds.: S.O. Schulze, C. Tschirner, R. Kaffenberger, S. Ackva)*, Carl Hanser Verlag, pages 115-124, 2017 [AMGS17].

- S. Melzer, U. Wittke, H. Hintze, and R. God. **Physische Architekturen variantengerecht aus Funktionalen Architekturen für Systeme (FAS) spezifizieren**, *Tagungsband zum Tag des Systems Engineering (Eds.: S.O. Schulze, C. Tschirner, R. Kaffenberger, S. Ackva)*, Carl Hanser Verlag, pages 429-438, 2016 [MWHG16].

- T. Bahns, S. Melzer, R. God, and D. Krause. **Ein modellbasiertes Vorgehen zur variantengerechten Entwicklung modularer Produktfamilien**, *Tagungsband zum Tag des Systems Engineering (Hrsg.: Chr. Muggeo, S.O. Schulze)*, Carl Hanser Verlag, pages 141-150, 2015 [BMGK15].

- S. Melzer, R. God, T. Kiehl, R. Möller, and M. Wessel. **Identifikation von Varianten durch Berechnung der semantischen Differenz von Modellen.** *Tagungsband zum Tag des Systems Engineering (Eds.: M. Maurer, S.O. Schulze)*, Carl Hanser Verlag, pages 279-288, 2014 [MGK$^+$14].

With the computation of A-box differences an extended form of an abduction problem will be solved. Approaches for solving abductive problems for multimedia data such as text and images are published here:

- S. Espinosa, A. Kaya, S. Melzer, and R. Möller. **On Ontology Based Abduction for Text Interpretation.** *Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics (Ed.: A. Gelbukh)*, number 4919 in LNCS, Springer, pages 194-205, 2008 [PKMM08].

- S. Espinosa, A. Kaya, S. Melzer, R. Möller, and M. Wessel. **Towards a Media Interpretation Framework for the Semantic Web.** *The 2007 IEEE/ WIC/ ACM International Conference on Web Intelligence (WI'07)*, IEEE Computer Society, Washington, DC, USA, pages 374-380, 2007 [PKM$^+$07b].

The following contribution presents a way for systematically combining symbolic (ontology-based) and holistic content descriptions in context of knowledge management in order to increase recall while at least maintaining precision.

- S. Melzer. **On the Relationship between Ontology-based and Holistic Representations in a Knowledge Management System.** *Ontology-based Applications for Enterprise Systems & Knowledge Management (Eds.: M. Nazir Ahmad, R. Colomb, and M. Abdullah).* IGI Global, pages 292-323, 2013 [Mel13].

More specifically, in the latter contribution an investigation for holistic content descriptions and retrieval of documents are based on the *latent semantic indexing* (LSI) [MRS08b, DDF$^+$90b], and symbolic (ontology-based) content descriptions are based on descriptions logics (DLs) (see Section 3.3.1).

1.4 Outline

This thesis is structured as follows: Chapter 2 introduces the fundamental aspects of knowledge management systems and the management of knowledge representation for documents. The characteristics of KM and the fundamentals of the representation of symbolic and holistic content descriptions via semantic assets are described in Chapter 3. The goal of Chapter 4 is to present current technologies and tools for representing content holistically and symbolically. General issues about content descriptions, which are essential to build semantic assets in context of KM as well as the latent semantic indexing approach for representing holistic content descriptions are presented. In addition, description logic as a formalism for the representation of symbolic content descriptions is presented. Chapter 5 describes the realization of semantic assets using the systematic combination of symbolic and holistic representations. After an evaluation of the systematic combination methodology, we conclude this thesis in Chapter 6 by summarizing the major contributions of this work and by presenting promising directions for future research.

Chapter 2

Content and Knowledge Management Characteristics: The Motivation for Semantic Assets

In the last years many research projects for supporting human work were carried out, emphasizing different aspects of the management of documents, images, videos, or persons: content management (CM), document management (DM), information management (IM), knowledge management - to name just a few [Seh04]. The general term content management also includes conceptual content management (CCM), web content management (WCM), and enterprise content management (ECM). Most of these terms are not defined succinctly, and indeed, all these management systems have the same purpose, namely the management of media data.

Data is represented in a media-specific form and in a structured, semi-structured, or unstructured way. This kind of data is considered as unstructured because it requires complex processes to allow computers to process the content behind it, or represented by it, beyond merely displaying the data on an output medium. We use the notion semi-structured for representations that contain symbolic annotations inside the data that allow for machine processing tasks to be supported by standardized notations such as RDFa or RDF Data

Cube [W3C14b]. In addition, data can be extended by so-called feature-based metadata based on standards such as XMP [Int12] (or, more specifically, Dublic Core). Interestingly, holistic representations of data are derived only for specific purposes (e.g., information retrieval), but are not intrinsically related to the data itself, and besides RDF Data Cube, which covers only statistics data, there are currently no standards available for systematically relating symbolic with holistic descriptions of media data.

Media data, i.e., structured, semi-structured, or unstructured data, are relevant for human information processing, and are therefore often referred to as *content*. The way in which content is handled by machines depends on the data model used, but data models are constrained by technical issues of the target system, for example imposed by databases. The notion of content tries to emphasize that media data should be handled at a more abstract layer, relevant for human problem solving tasks defined at a conceptual level. A conceptual model can avoid dealing with low-level technical constraints such as details of combining media data with metadata, say, as shown in [SS04]. Schmidt and Sehring introduce *assets* in order to specify such a conceptual model. The asset definition is presented in Section 2.1. In a knowledge management scenario, content is organized and presented in such a way that by "consuming" content, humans can increase their knowledge. This holds in particular in industrial knowledge management settings. Knowledge is used in a context-specific way and depends on the situation at hand [DACN03, HH15, HK15]. Consequently, it is essential to know how to create knowledge on the one hand, and how to handle knowledge on the other in order to support productivity increases. While speaking at the knowledge level might be appropriate for planning purposes, eventually knowledge management needs to be done at the content level and, lastly, at the data level.

The main purpose of this section is to give an overview about the characteristics of conceptual content and knowledge management. A detailed analysis of KM and CCM notions that are essential for the formalization of KM notions via semantic assets is described in Section 5.

2.1 Conceptual Content Management

Content management (CM) applies a set of processes and technologies for creating and managing content, and supports the evolutionary life cycle of content (texts, diagrams, codes, data, and so on). A content management system (CMS) is a tool that enables a variety of centralized technical and non-technical staff to create, edit, and publish various forms of content. Content is managed by a set of rules, processes, and workflows such that CM systems ensure coherent and validated representations of data. Moreover, a CMS must enable users to collaborate and interact for the creation and management of trusted content through a so-called portal, and also must allow users to import new content [SCRP09]. In particular the import functionality makes clear that media data, feature-based metadata, and symbolic content descriptions need to be combined under a single “handle”. In order to combine both aspects in a concrete situation, Schmidt and Sehring coined the notion of an *asset* [SS03]. According to Schmidt and Sehring, assets are used to provide descriptions of entities through pairs of media content and conceptual abstractions in a CCM system (see Figure 2.1) [Seh04, SS04, SS03].

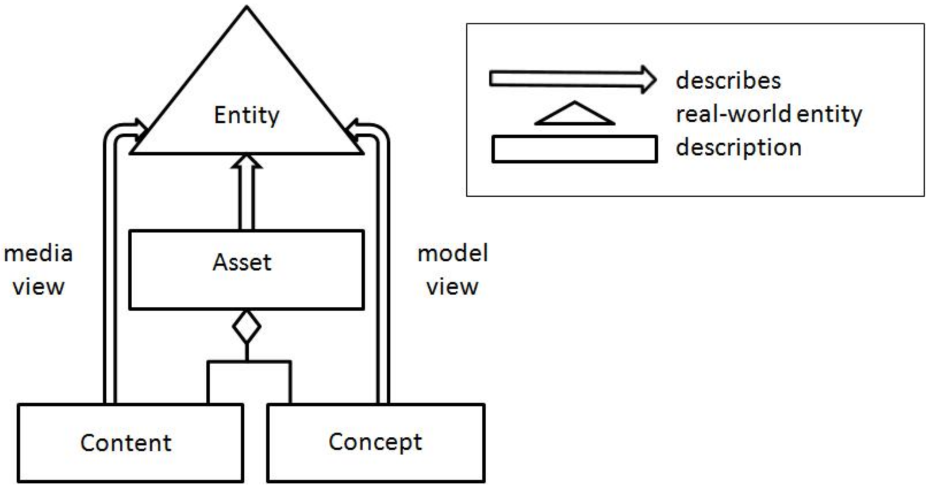


Figure 2.1: **Dualistic description of entities by assets.** Adapted from [Seh04].

The content part is a reference to the raw media data (media view) and the concept part contains feature-based metadata for describing media data and a single entity being modeled.

Interestingly, in the [SS04] approach, there are no holistic representations of content considered. In addition, the concept part considers just one entity, and is therefore in some sense a local representation. For knowledge management in the large, we argue that assets need to be extended with holistic representations, which, due to their nature, provide an integrated view on a whole repository of entities, albeit there is a holistic representation still associated with a particular media object. The intricate relationship of local (media-specific) representations and global representations (influences from repository context) is captured by our extension of Sehring's and Schmidt's assets [SS04] to so-called *semantic assets*.[1]

2.2 Knowledge Management

In most companies a lot of knowledge is available, but management often does not acknowledge the importance of this knowledge [SMD12]. Many companies do not work with existing knowledge because they are not aware of the knowledge they already possess. Therefore, a lot of knowledge remains unused, and therefore according to [SMD12] the core of "knowledge management (KM) is about trying to harvest all the insights and experience that go into making an organization function". To this end, knowledge management can provide a more effective and efficient usage of knowledge. The idea is that knowledge of an employee is made available for other employees in order to ensure the long-term success of the company.

In most of the literature about KM, knowledge is divided into *explicit* knowledge and *implicit* knowledge (also called tacit knowledge).

Explicit knowledge Explicit knowledge is knowledge that can be captured and written down in documents.

Implicit knowledge Implicit knowledge refers to the skill that people possess, and skill is hard to communicate.

[1] Note that the terms content and concepts described in the works of Schmidt and Sehring are not exactly the same as used in this work. Schmidt and Sehring define concept as a description for concrete and abstract entities.

In [SS00] Schneider argues that it is important to share and to manage explicit knowledge as well as implicit knowledge because both types of knowledge are essential for the development of enterprises. This opinion is confirmed by Nonaka [Non08] and others. Not only is the consideration of explicit and implicit knowledge essential, but also the management of this knowledge. In 1986 Wiig defined the term *knowledge management* in [Wii86]. He proclaims that the main objectives of a knowledge management system (KMS) are:

- Making the enterprise act as intelligently as possible to secure its viability and overall success.
- Realize the best value of its knowledge assets.

There are two fundamental aspects of KM. The first aspect involves knowledge being considered as an asset that is capable of being shared within a wider community. The second aspect considers that there should be a balance between explicit and implicit knowledge [CW99]. Many early knowledge management projects involve intranet solutions to keep and distribute a form of "knowledge" inside companies. The visibility of the asset model enabled its use to justify significant levels of investment, principally in technology-based solutions.

In [DB05] KM is defined as follows: KM is based on the idea that an organization's most valuable resource is knowledge of its people. Therefore, the extent to which an organization performs well will depend, among other things, on how effectively its people can create new knowledge, share this knowledge within the organization, and use that knowledge most effectively. Schütt describes in [Sch03] two generations of knowledge management.

Generation 1 [Sch03] The need for knowledge management strategies was identified, in combination with a new role: the Chief Knowledge Officer (CKO), at best directly reporting to the CEO. CKOs immediately went to the first KM conferences in the UK or the US. They all only had one question in their mind: "What is my job?" Unfortunately, they did not received an answer and so most of them resigned from their positions within a year. That was Generation 1 of Knowledge Management, for early adopters roughly lasting from 1990 to 1995.

Generation 2 [Sch03] From 1995 to 2000: the erroneous belief that knowledge can be codified to a large extent came into managers' minds. The theoretical background for it had been created by Nonaka. He had published some work on information creation and since 1991 he used the label "knowledge creation". His article "The Knowledge Creating Company" in Harvard Business Review brought some attention, but the real breakthrough had to wait until 1995, when he published a book with the same title [NT95], together with Takeuchi. In this book a knowledge creation process, called SECI process, is described. The SECI process defines "WHAT" kind of knowledge is required. In subsequent contributions of Nonaka et al. (cf. [NK98], and [NTB03]), the SECI process is extended.

Generation 3 Knowledge representation in the third generation gets a new meaning by considering semi-automated knowledge management processes. Such a partially automated knowledge management process is presented in this thesis by describing the systematic combination of holistic and symbolic representations and the integration of this method in a knowledge creation process (see Section 5.3). Documents produced in previous project a company carried out automatically contribute to holistic representations, as will be explained below. In total, third generation approaches define "WHAT" kind of knowledge is required, "HOW" each task is performed, and which tools can be used.

2.2.1 Knowledge Creation Model

The challenge of early knowledge management approaches was to develop models to make implicit knowledge explicit while allowing for explicit knowledge to be made individually meaningful.

In [NTB03] Nonaka, Toyama, and Byosière propose a multilayered process of knowledge creation in order to understand how companies create knowledge dynamically and how knowledge might be actually transferred. For that to happen, knowledge has to be "transformed to information", and only then can it be "moved" [SMD12].

The Nonaka approach of knowledge creation is based on the SECI process, a platform for knowledge creation, and “knowledge assets”, which are the inputs and outputs of the knowledge creation process. We revisit each of these notions in the following paragraphs.

SECI process

Nonaka et al. sought to establish a sense of sharing of explicit and implicit knowledge in the knowledge transfer model, and to this end they proposed the SECI model which has four sub-processes:

- Implicit to implicit: socialization (S),
- Implicit to explicit: externalization (E),
- Explicit to explicit: combination (C), and
- Explicit to implicit: internalization (I).

Socialization Socialization is the process that transfers implicit knowledge of one person to implicit knowledge of another person [Non08]. According to [SMD12] this process can be seen as an experiential, active and a “living thing” process.

Externalization Externalization is the process for making implicit knowledge explicit [Non08]. Or, as [SMD12] characterize it, with externalization, implicit knowledge is translated into a readily understandable form (explicit knowledge).

Combination Combination is the process for transferring one’s explicit knowledge into explicit knowledge [Non08]. Combination provides for an increased usability of explicit knowledge. Information technology is readily suitable for realizing such a process because explicit knowledge can be conveyed in documents, email, or databases [SMD12].

Internalization Internalization is the process of understanding and absorbing explicit knowledge into implicit knowledge [SMD12, Non08]. For [SMD12], implicit knowledge is “executable” by the owner. However, internalization is largely experiential.

The SECI process defines "WHAT" kind of knowledge is required for KM. In addition, Nonaka and Konno define an environment for knowledge sharing processes called *ba*.

***Ba*: The environment for knowledge creation**

The creation of new knowledge requires a environment for sharing, creating and utilizing knowledge. The *ba* concept [NK98, NT01] defined by Nonaka and Konno offers such an environment. They have extended the SECI model with four types of *ba* environments. The four types correspond to the four modes of the SECI model:

- Socialization with originating *ba*: The originating *ba* is a place where individuals share feelings, emotions, experiences, and mental models.

- Externalization with dialoguing *ba*: The dialoguing *ba* is a place where selected people with a specific knowledge interact during a face-to-face communication with other people with a similar specific knowledge. For example, mental models from the originating *ba* are shared through concepts, articulation of their thinking, and so on.

- Combination with systemizing *ba*: The systemizing *ba* is a place where new explicit knowledge is combined with other explicit knowledge. This kind of knowledge transfer is accomplished among groups across organizations [SMD12].

- Internalization with exercising *ba*: The exercising *Ba* is a place where the conversion of organization and group explicit knowledge to the individual implicit knowledge is facilitated [SMD12].

The *ba* concept offers a knowledge transfer environment and defines the "WHERE" of each SECI mode, but neither the "HOW" nor the "WITH WHAT". The "WITH WHAT" question is answered by Nonaka et al. with the definition of so-called *knowledge assets*.

Knowledge assets

In the context of KM inspired by Nonaka, so-called *knowledge assets* are the basis for knowledge creation. In [NTB03], Nonaka, Toyama, and Byosière define knowledge assets as follows: Nonaka's knowledge assets are inputs and outputs of the knowledge creation process. They are unseizable and have a limited lifetime in contrast to physical assets. In [NT01], Nonaka and Teece divide knowledge assets into four types corresponding to the SECI modes:

- Experiential: implicit knowledge shared through common experiences
- Conceptual: explicit knowledge articulated through images, symbols, and language
- Systemic: systematized and packaged explicit knowledge
- Routine: implicit knowledge embedded in actions and practices

Knowledge assets cannot easily be managed in the traditional way of management (for instance with CCM) because they change over time. In [NKT98] Nonaka, Konno, and Toyama present a new method about "HOW" knowledge assets can be dynamically managed, and they define a creation process for knowledge assets. Figure 2.2 illustrates this knowledge creation process with the three basic concepts SECI, *ba*, and knowledge assets (KA).

A company uses existing knowledge assets and creates new knowledge through the SECI process taking place in the *ba* environment. Newly created assets will be added to the existing knowledge assets of the company. Management of these assets encompasses the following activities: All users can work on all three elements of the knowledge creation process: build and set up (energize) the tool environment (*ba*), execute or lead the SECI process, and define the *knowledge vision*. The knowledge vision defines what kind of knowledge the company should create in what a domain. The definition of a knowledge vision supports the realization of dynamic knowledge management because it gives a direction where the company should be going and defines how knowledge can be managed over a long-term period [DACN03].

Nonaka's model of knowledge management played a crucial role in understanding "HOW" to create and share knowledge in general. However, there

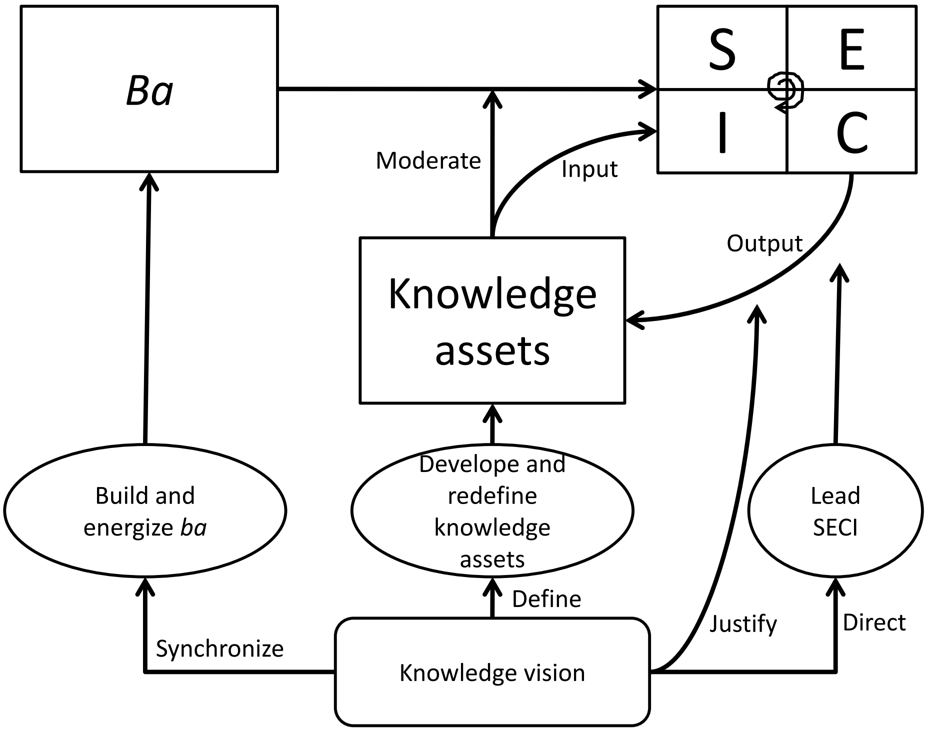


Figure 2.2: **Knowledge creation process**. Description see text. Source [NKT98].

are fundamental problems with this model because it is too informal, and it lacks nuance and sophistication to be made useful across different companies, countries, and over time [ES97, SMD12]. This is also confirmed by some case studies mentioned in [SMD12]. In addition, the authors in [SMD12] mention that knowledge management, conversion, and codifying requires further research and development to take into consideration the implicit origins (set of documents) of knowledge and the rapidly changing methods of communication (provision and exchange of documents).

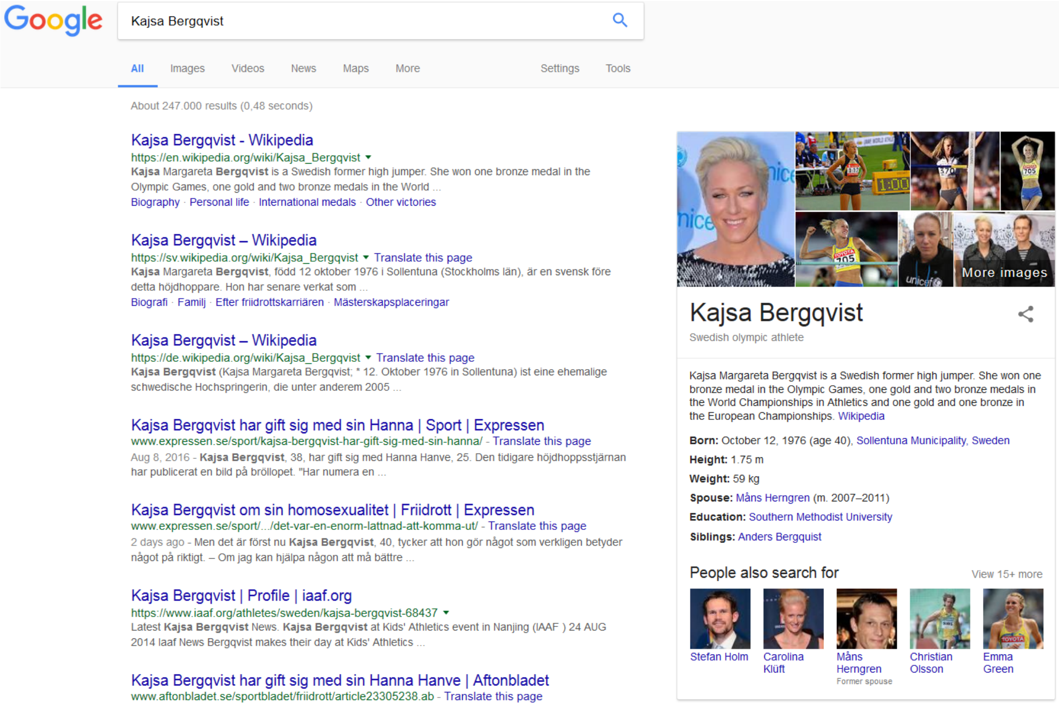
We argue and investigate in this thesis that Nonaka's visionary model is indeed practically realizable by (i) defining concrete processes for all four SECI modes, (ii) using Nonaka's method for knowledge creation, and (iii) by defining concrete knowledge management units. Nonaka's knowledge assets are implemented in this thesis as semantic assets with feature-based as well as holistic and symbolic representations. The formalization essentially relies on systematically combining holistic and symbolic content descriptions, and is this basis for information retrieval tasks supporting knowledge creation work as part of the tools being used in the ba. We explain the formalization in detail later with a concrete knowledge management scenario, which is explained in the next subsection.

2.2.2 A Knowledge Management Scenario

Knowledge Management is important for large and challenging engineering projects. Obviously, in this thesis we cannot deal with fully-fledged engineering problems. Nevertheless, for illustration purposes we roughly consider the concept of building a sports stadium. In 1966, the Munich Olympic Stadium was built that meets the concept "Green Olympic Games" [Wik15a]. The number of disciplines engineers have to know about increase while designing, for example, an olympic stadium. Besides science and engineering-oriented challenges, engineers require knowledge about what is important to fulfill a concept such as "Green" for building a sports stadium. We assume that, first of all, in general an engineer starts to search for information about the concept of, say, a green stadium, using information retrieval (IR) systems. Second, while starting a construction, engineers could also benefit from additional information being made available to them automatically and in a proactive way. Thus, their knowledge is extended due to meet the challenges involved in a very specific construction task without requiring them to pose particular queries. This kind of ba could increase productivity and creativity of engineers, and this is what we have in mind with knowledge management based on semantic assets.

For constructing the athletic areas, for instance jumping areas, in a stadium, engineers might be inspired by having at look at descriptions of athletic events or particular athletes. Using standard information retrieval with, for instance, the name of a concrete athlete such as "Kajsa Bergqvist" the user is shown references to documents (web pages, Youtube videos, etc.) as well as symbolic descriptions of the person (Google, Knowledge Vault [DMG^{+}14], see Figure 2.3).

While Google's Knowledge Vault (KV) enables certain kinds of specific follow-up queries, one can, for example, not easily search for other high jumpers in an olympic context. Selecting "high jumper" and querying Google leads to



Kajsa Bergqvist

All Images Videos News Maps More Settings Tools

About 247.000 results (0,48 seconds)

Kajsa Bergqvist - Wikipedia
https://en.wikipedia.org/wiki/Kajsa_Bergqvist ▾
Kajsa Margareta **Bergqvist** is a Swedish former high jumper. She won one bronze medal in the Olympic Games, one gold and two bronze medals in the World ...
Biography · Personal life · International medals · Other victories

Kajsa Bergqvist – Wikipedia
https://sv.wikipedia.org/wiki/Kajsa_Bergqvist ▾ Translate this page
Kajsa Margareta **Bergqvist**, född 12 oktober 1976 i Sollentuna (Stockholms län), är en svensk före detta höjdhoppare. Hon har senare verkat som ...
Biografi · Familj · Efter friidrottskarriären · Mästerskapsplaceringar

Kajsa Bergqvist – Wikipedia
https://de.wikipedia.org/wiki/Kajsa_Bergqvist ▾ Translate this page
Kajsa Bergqvist (Kajsa Margareta Bergqvist; * 12. Oktober 1976 in Sollentuna) ist eine ehemalige schwedische Hochspringerin, die unter anderem 2005 ...

Kajsa Bergqvist har gift sig med sin Hanna | Sport | Expressen
www.expressen.se/sport/kajsa-bergqvist-har-gift-sig-med-sin-hanna/ - Translate this page
Aug 8, 2016 - **Kajsa Bergqvist**, 38, har gift sig med Hanna Hanve, 25. Den tidigare höjdhoppsstjärnan har publicerat en bild på bröllopet. "Har numera en ...

Kajsa Bergqvist om sin homosexualitet | Friidrott | Expressen
www.expressen.se/sport/.../det-var-en-enorm-lattnad-att-komma-ut/ - Translate this page
2 days ago - Men det är först nu **Kajsa Bergqvist**, 40, tycker att hon gör något som verkligen betyder något på riktigt. – Om jag kan hjälpa någon att må bättre ...

Kajsa Bergqvist | Profile | iaaf.org
https://www.iaaf.org/athletes/sweden/kajsa-bergqvist-68437 ▾
Latest **Kajsa Bergqvist** News. **Kajsa Bergqvist** at Kids' Athletics event in Nanjing (IAAF) 24 AUG 2014 Iaaf News Bergqvist makes their day at Kids' Athletics ...

Kajsa Bergqvist har gift sig med sin Hanna Hanve | Aftonbladet
www.aftonbladet.se/sportbladet/friidrott/article23305238.ab - Translate this page

More images

Kajsa Bergqvist
Swedish olympic athlete

Kajsa Margareta Bergqvist is a Swedish former high jumper. She won one bronze medal in the Olympic Games, one gold and two bronze medals in the World Championships in Athletics and one gold and one bronze in the European Championships. Wikipedia

Born: October 12, 1976 (age 40), Sollentuna Municipality, Sweden

Height: 1.75 m

Weight: 59 kg

Spouse: Måns Herngren (m. 2007–2011)

Education: Southern Methodist University

Siblings: Anders Bergquist

People also search for View 15+ more

Stefan Holm · Carolina Klüft · Måns Herngren (Former spouse) · Christian Olsson · Emma Green

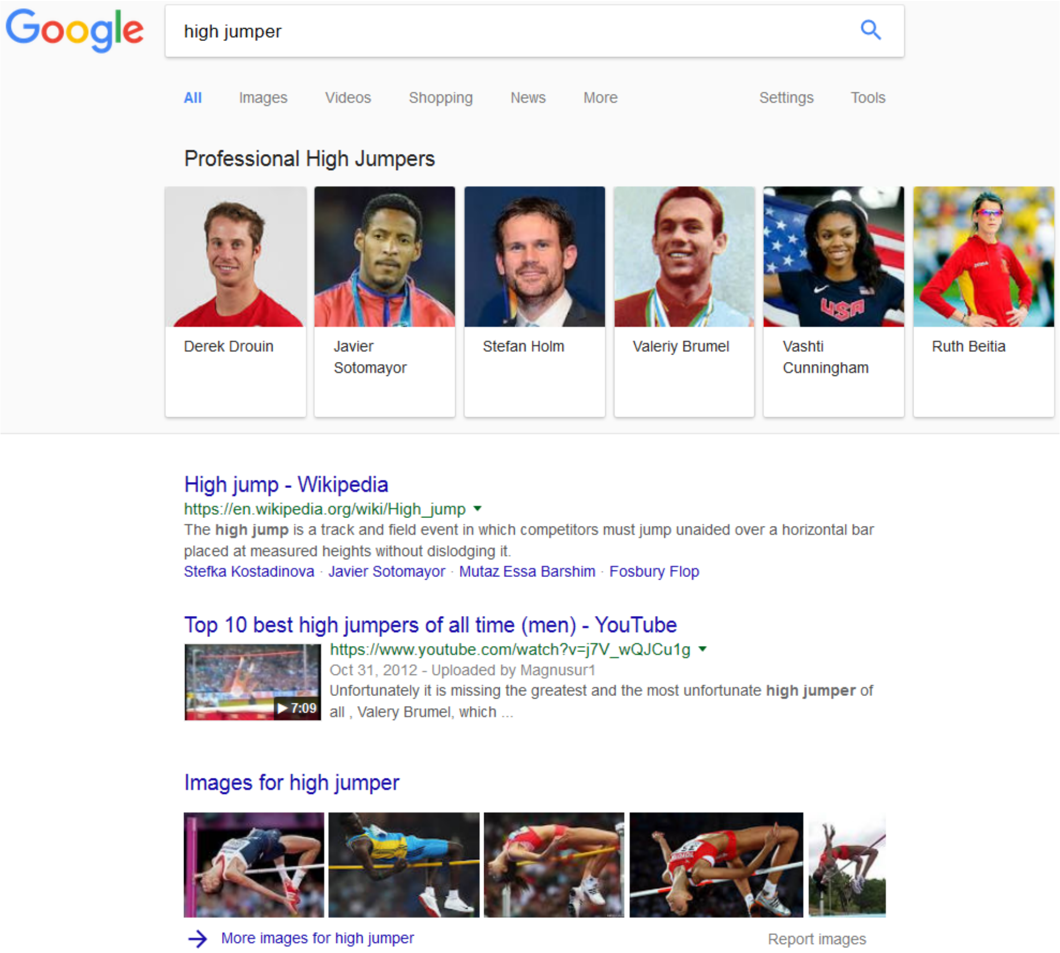
Figure 2.3: **Google information retrieval results for the string query “Kajsa Bergqvist” are, on the left-hand side, a list of references to documents, and, on the right-hand side, symbolic descriptions of the person.**

unspecific answers concerning various kinds of high jump athletes with high recall (see Figure 2.4).

However, Google does not allow for easily increasing precision by focusing on Kajsa Bergqvist high jump events apparently because the text “high jumper” is not considered which is below the images in Figure 2.3 at the right-hand side.

Previously, engineers had to search manually for “jumping events” in order to gain information that “hurdling”, “long jump”, and “pole vault” belong to “jumping events”. Engineers had to repeat this manual process for each jumping event type in order to obtain particular references to documents or symbolic representations because of the missing latent structure of jumping events. The latent structure of all documents can be retrieved if the data of linked documents associate with one another (see Figure 2.5, left).

In order to efficiently deliver relevant document links to the user, the challenge of IR systems is to filter, prioritize, and efficiently deliver results to users



Professional High Jumpers

Derek Drouin · Javier Sotomayor · Stefan Holm · Valeriy Brumel · Vashti Cunningham · Ruth Beitia

High jump - Wikipedia
https://en.wikipedia.org/wiki/High_jump
The **high jump** is a track and field event in which competitors must jump unaided over a horizontal bar placed at measured heights without dislodging it.
Stefka Kostadinova · Javier Sotomayor · Mutaz Essa Barshim · Fosbury Flop

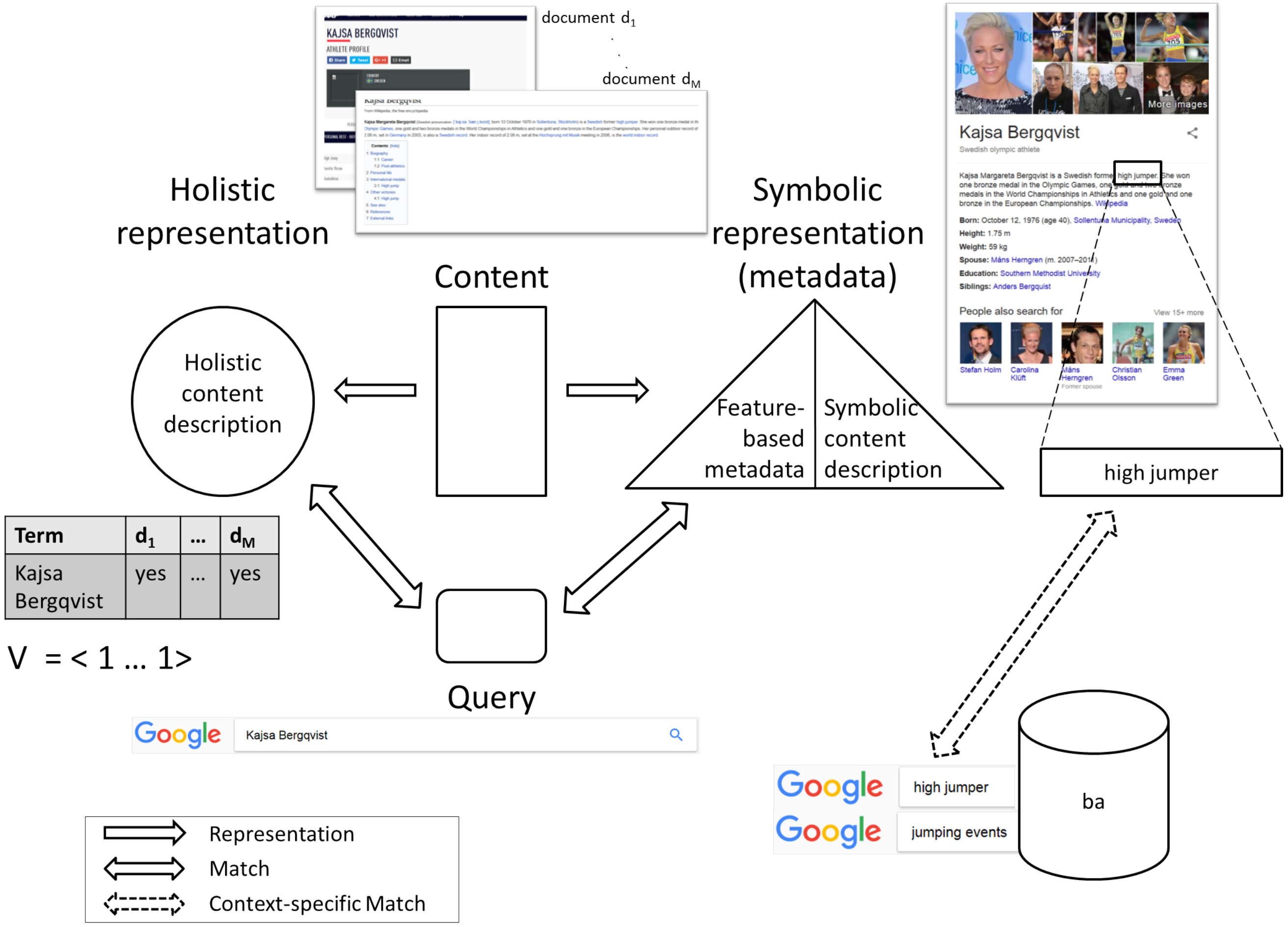
Top 10 best high jumpers of all time (men) - YouTube
https://www.youtube.com/watch?v=j7V_wQJCu1g
Oct 31, 2012 - Uploaded by Magnusur1
Unfortunately it is missing the greatest and the most unfortunate **high jumper** of all , Valery Brumel, which ...

Images for high jumper

Figure 2.4: **Google information retrieval results for the string query "high jumper" are, at the top as a list of references to symbolic descriptions of a person, and, at the bottom as a list of references to documents.**

with high recall and high precision. For our knowledge management scenario we suggest to consider the engineer's working context in the ba environment so that engineers can easily search for related information and the IR system acts in a proactive way.

Figure 2.5 illustrates that holistic and symbolic representations can be used in order to find relevant documents for the string query "Kajsa Bergqvist" using context-specific data from the ba environment. The idea is that each document $d_1 \ldots d_M$ has symbolic and holistic representations. In this example, the symbolic representation is delivered by Google's KV (see Figure 2.5 on the right-hand side). The holistic representation, here presented as a simplified example, is a vector $\vec{V}$ which represents document representations with relations to other documents. If the term 'Kajsa Bergqvist" occurs, in $\vec{V}$ the value is 1, and 0 otherwise (see Figure 2.5 on the left-hand side). One can imagine a ba environment as a local repository with previous posed queries, here "high jumper" and "jumping events." The aim is to use the ba so that an IR system can act in a proactive way. It follows that the term "high jumper" in the text of the symbolic description is highlighted and has a new link to high jumper content. In order to achieve these new links, the challenge is to systematically combining holistic and symbolic information retrieval approaches. A systematic combination approach is presented in Chapter 5. For implementation purposes, we need formal preliminaries introduced in the next two chapters.



Term	d_1	...	d_M
Kajsa Bergqvist	yes	...	yes

V = < 1 ... 1>

Figure 2.5: **An illustration of systematically combining holistic and symbolic IR approaches in a ba environment** (for a description see text).

Chapter 3

Fundamentals for Semantic Assets

The discussion of the previous knowledge management scenario and several other research contributions [Ing99, DSdGM15, HK15] shows that there is a need for linking information retrieval results with context-specific data for supporting the creativity of engineers. The aim of this thesis is to link semantic and context-specific annotations in a systematic way, such that the latent structure of a specific domain an engineer is interested in is addressed automatically. We build on the idea that a formalization for the ba environment, as well as Nonaka's overall knowledge creation process, provides support for making available additional information automatically in a proactive way. Before we present our methodology for knowledge management using semantic assets in Chapters 4 and 5, we give some formal preliminaries.

3.1 Representation of Content Descriptions

We now explain the symbolic notions of content interpretation, as defined by philosophers, because these notions play a fundamental role in the area of content management and content descriptions [SS03, Bos08]. In addition, we outline holistic content descriptions, which could be used for representing context-specific data.

The notion of a symbol is defined by Cassirer [CN64]. He defines a symbol as an invisible unit capable to encompass the totality of phenomena in which the sensuous symbol is in any way filled with meaning, following from a process of symbolic formation [Cas23]. In other words, Panofsky defines a symbol as a synthetic intuition that could identify the proper meaning of the content [Pan75]. The way how content becomes a symbol is already defined in [Pan70]. More concretely, Panofsky develops a methodological approach to the systematic specification of objects, especially objects of art. He distinguishes three levels of description (cf. [Sch09]):

Pre-iconographical level (level 1) The pre-iconographical level is a description level for objects found in document content, and represents the particular characteristics of objects, such as being a person, horizontal bar, or medal.

Iconographical level (level 2) The iconographical level gives the meaning of objects by introducing the specific iconographic vocabulary which could be used in a context-specific way. At this level, person, horizontal bar, and medal, as defined at the pre-iconographical level, are now specialized to athlete, crossbar, and honorary award, respectively.

Iconological level (level 3) The iconological level additionally represents objects by general cultural effects. At this level, athlete, crossbar, and a honorary award, for example, together represent a high jump champion.

According to Panofsky, symbols are classification labels generated from the object content, and symbol generation can be framed as a problem of classifying an image region to one of several objects [MLLK04]. The technical implementation of Panofsky's approach is carried out by traditional databases (level 1), (digital) libraries (level 2), and web-provided content (level 3). However, Panofsky's methodology to image description and understanding was not unanimously accepted because, mentioning just one reason for criticism, there was no room for any elements of stile and form used by an artist [Sch09]. In [SS03] and [Bos08] the authors address the issue and define *assets* [SS03] as a structurally rich way to capture explanations of content. However these works show that the management of content with their meaning is not easily accomplished. The reason is that “symbols”, as we see then in this work, need to be sets of first-order logic formulas (see Section 3.3) and not simple classification labels. In this thesis, these symbolic content descriptions are managed with so-called *semantic assets*.

The approaches presented in [DMG$^+$14] are used to represent information via symbolic descriptions, e.g., taken from Wikipedia[1], derived by Open Calais[2], and described with RDFa (Resource Description Framework in Attributes) [W3C14a]; but the relation to a cultural environment is missing, which is essential, as we presented in the knowledge management scenario, for providing support for the work of an engineer. In the following we describe approaches for representing content holistically and symbolically in combination with their pros and cons (Sections 3.2 and 3.3). Based on both approaches we present a new methodology in Section 5.1 and a methodology in which context-specific data can be anchored (Section 5.3).

3.2 Holistic Representations

The most popular holistic approach for representing document content is the *TF-IDF* scheme [Jon72, SM86], in which a basic vocabulary of M terms is chosen, and, for each of the N documents in the repository, a count is formed as the number of occurrences of each term in a document. After normalization, a term count (“frequency”) is related to an inverse document frequency count, which measures the number of occurrences of a term in the entire corpus. The result is an $M \times N$ term-document matrix whose columns contain the IDF values for each of the documents in the repository. Thus, documents of arbitrary length are reduced to vectors of M numbers [BNJ03], considering other documents in the repository to some extent in the inverse document frequency measure. A semantic space is built wherein similar documents are located closed to one another. In addition, semantic indexing uses the terms in a query to identify a pseudo document as a point in the semantic space. The central idea exploited for information retrieval (IR) is that documents in the neighborhood of this point are returned to the user [SWY75].

[1] `https://en.wikipedia.org/`

[2] `http://www.opencalais.com/`

3.2.1 Latent Semantic Indexing

Given the TF-IDF matrix, the idea of latent semantic indexing (LSI) is to implement a dimension reduction procedure for deriving a structure "hidden" behind the TF-IDF numbers [DDF$^+$90b]. Formally, the $M \times N$ matrix TF-IDF is defined as

$$\text{TF-IDF}_{i,j} := \text{TF}(t_i, d_j) \cdot \text{IDF}(t_i) = \frac{\#terms(t_i, d_j)}{M}(1 - \log \frac{N}{\#docs(t_i)}), \quad (3.1)$$

where $\#terms(t_i, d_j)$ denotes the number of occurrences of term t_i in document d_j, and $\#docs(t_i)$ denotes the number of documents in the whole repository in which the term t_i appears. We consider only terms that appear somewhere in the repository.

The hidden structure is exploited to deal with synonymy and polysemy effects to better support IR. An M-dimensional document vector is approximated by a vector of k numbers ($k < M$), and so is a query pseudo document. In this new space similarity-based retrieval produces higher recall, thus coping with synonymy and polysemy effects w.r.t. terms used in the repository. The LSI approach is described in detail below.

Fundamentals of latent semantic indexing

Let C be an $M \times N$ matrix, where the $M \times N$ matrix is a term-by-document matrix with non-negative values. Each row corresponds to a unique term in the document corpus and each column corresponds to a document.

For the rank of C, an $M \times N$ matrix, it holds that $rank(C) \leq min\{M, N\}$. Furthermore, a square matrix can be a diagonal matrix, denoted $diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$, with the dimension $r \times r$. If all diagonal entries of a diagonal matrix are 1, this matrix will be called the identity matrix (of dimension r). The matrix is represented as

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \quad (3.2)$$

An orthogonal matrix C has the characteristics $C^TC = I$. For an $M \times M$ matrix C and a vector $\vec{x}$, the values of λ satisfying the equation

$$C\vec{x} = \lambda\vec{x} \text{ with } \vec{x} \neq 0, \tag{3.3}$$

are called eigenvalues of C. The associated vectors $\vec{x}$ are called (right) eigenvectors. This equation is equivalent to the following equation

$$(C - \lambda I_M)\vec{x} = 0 \text{ with } \vec{x} \neq 0. \tag{3.4}$$

There can be at most $r \leq rank(C)$ eigenvalues.

Singular value decomposition

Given an $M \times N$ matrix C, let U be the $M \times r$ matrix whose columns are the eigenvectors of CC^T, and V be the $r \times N$ matrix whose columns are the eigenvectors of C^TC. The *singular value decomposition* (SVD) for C is defined as

$$C := U\Sigma V^T, \text{ where} \tag{3.5}$$

$$U = \begin{pmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,r-1} & u_{1,r} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,r-1} & u_{2,r} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{m-1,1} & u_{m-1,2} & \cdots & u_{m-1,r-1} & u_{m-1,r} \\ u_{m,1} & u_{m,2} & \cdots & u_{m,r-1} & u_{m,r} \end{pmatrix} \tag{3.6}$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{r-1} & 0 \\ 0 & 0 & \cdots & 0 & \sigma_r \end{pmatrix} \tag{3.7}$$

$$V^T = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n-1} & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{r-1,1} & v_{r-1,2} & \cdots & v_{r-1,n-1} & v_{r-1,n} \\ v_{r,1} & v_{r,2} & \cdots & v_{r,n-1} & v_{r,n} \end{pmatrix} \tag{3.8}$$

For the LSI application we can assume $r = M$ because there are no duplicate documents in the repository. The idea of an LSI is to compute a rank-k-approximation C_k of C with low error, where $k < r$. It has been shown [EY36] that this can be achieved by setting $\sigma_{k+1} \ldots \sigma_r$ to zero.

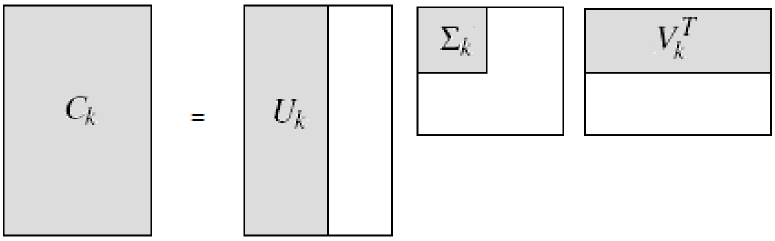


Figure 3.1: **Diagram of an SVD for a rank-k-approximation of low error.** Description see text.

The computation of C_k captures the important underlying semantic structure of types and documents [MB07] as shown in Figure 3.1. “The semantic structure is only the correlation structure in the way in which individuals words appear in documents; semantic implies the fact that terms in a document may be taken as referents to the document itself or to its topic“ [DDF$^+$90a]. The white areas indicate matrix entries that are zero. Σ_k contains k non-zero singular values. As can be seen $U\Sigma = U_k\Sigma_k$, where U_k is a projection of U to the first k columns. Analogously for V^T and V_k^T. Hence, a rank-k-approximation of low error is defined as

$$C_k := U_k \Sigma_k V_k^T. \tag{3.9}$$

The rank of C_k, an $M \times N$ matrix, is at most k. This follows from the fact that Σ_k has a most k non-zero values (see Figure 3.1).

The Frobenius norm of a matrix X is defined as

$$\|X\|_F := \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} X_{ij}^2}. \tag{3.10}$$

The matrix C_k is the best rank-k approximation of the original matrix C because the distance between C and C_k is minimized according to the Frobenius norm [EY36], i.e.

$$C_k := \arg\min_{Z|rank(Z)=k} \|C - Z\|_F \, . \tag{3.11}$$

Retrieval: Latent semantic indexing

A low-rank approximation of C yields a new representation for the set of documents in a repository. As we will see in the following, queries can been seen as documents and can also be represented using the low-rank approximation. Given such a query (document), similarities between query and document can be computed in the low-rank space. In this context the process of computing query-document similarity scores is known as latent semantic indexing.

In the latent semantic indexing process the value k is generally chosen in the low hundreds [MRS08b]. Thus k is far smaller than the original rank of C. Originating from Equation 3.9, we derive the holistic repository representation H as follows

$$H := V_k^T. \tag{3.12}$$

A string query is represented by a query vector $\vec{q}$.

$$\vec{q} := \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_M \end{pmatrix}, \tag{3.13}$$

where the values $q_1 \dots q_M$ are either 0 or 1. If a string is equal to a term, the value is 1, and 0 otherwise. The vector $\vec{q}$ is mapped into its representation in the LSI space via the following equation

$$\vec{q_k} := \Sigma_k^{-1} U_k^T \vec{q}. \tag{3.14}$$

The *cosine similarity* computes the distance between a query and a document, or between two documents. The cosine similarity (sim) between two documents d_1 and d_2 is defined as:

$$sim(d_1, d_2) := cos(\alpha) = \frac{\vec{d_1}\vec{d_2}}{\left\|\vec{d_1}\right\| \left\|\vec{d_2}\right\|}, \tag{3.15}$$

where α is the angle between the document vectors $\vec{d_1}$ and $\vec{d_2}$.

Given the similarities between a query $\vec{q_k}$ and column vectors $\vec{d}$ from H (document vectors), the documents with similarity values beyond a given threshold are selected as a query result.

3.2.2 Related Approaches

In the field of IR, many researchers optimize LSI in order to increase recall and precision. While the *TF-IDF* provides a relatively small amount of reduction of document representations or structure, LSI addresses these shortcomings. However, further optimization approaches which are based on LSI are developed such as *probabilistic latent semantic analysis* (PLSA) [Hof99] and *Latent dirichlet allocation* (LDA) [BNJ03].

PLSA is a method for unsupervised learning, which is based on a statistical latent class model. This class model is called *aspect model* which is a latent variable model containing observations in the form of co-occurrences of words and documents $(w|d)$. PLSA computes the probability of each co-occurrences as a mixture of conditionally independent variables, formally: $P(w|d) = \sum_{z \in Z} P(z)P(d|z)P(w|z)$, where the variable z is a observed class variable $z \in Z := \{z_1, \ldots, z_p\}$, i.e., the words' topic [Hof99]. A further optimization method for PLSA is presented in [Hof03], in which a novel statistical class model is used for IR tasks. The new approach is called Gaussian pLSA.

LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Further on, each topic is modeled as an infinite mixture over an underlying set of topic probabilities. These topics represents the implicit knowledge behind a document. LDA can be viewed as a dimensionality reduction technique, in the spirit of LSI with proper underlying generative probabilistic semantics [BNJ03]. In other words we see LDA is an optimized LSI method. However, these approaches demonstrate that LSI is a very good basis for IR tasks because such systems delivers documents with high recall. But precision decreases when recall increases (cf. experiments in [KFN09]).

LSI is one implemented model for distributional semantics and varies w.r.t. the usage of i.e., dimension reduction or similarity measure. With regard to the performance of cosine similarity measure, Locality Sensitive Hashing (LSH) [IM98] is a more efficient approach. The basic idea of LSH is to hash objects into the same bucket in such a way that the probability of collision is much higher for objects that are close to each other than for those which are far apart. In the LSH method, there is a possibility for the emergence of false positive and false negative. False positives are dissimilar objects which are hashed to the same bucket, and false negatives are similar objects which are not dispatched to the same bucket [AM13]. In [AM13] it is shown how to reduce false positives and false negatives.

In this thesis we present a novel approach for finding latent structures for knowledge management which is based also on LSI in order to receive IR results with high recall. For increasing precision a symbolic approach is used for document representations. In a systematic combination of both methods, we present how to retrieve IR results with high recall and at least maintaining precision.

3.3 Symbolic Representations

Nowadays, many documents in repositories contain not only textual but also visual and auditory information [Kay11]. Despite this fact, retrieval techniques that rely only on information from textual sources (i.e., surrounding texts of web sites) are still widely used due to the success of existing IR software systems, in particular with respect to stability and scalability [DMG$^+$14]. Despite the idea that the right part of the web page in Figure 2.3 in Section 2.2.2 should indeed support follow-up queries, this holds only for certain items. Google's technique does not allow for easily increasing precision by focusing on Kajsa Bergqvist as a high jumper (see the text that is presented for describing Kajsa Bergqvist), which would be interesting in our knowledge management scenario from above. The user is forced to type in another query with “Kajsa Bergvist high jumper” as a query text. The system is not able to compute follow-up query possibilities on the fly. This would require the consideration of the context (the ba in terms of knowledge management) as well as symbolic representations (annotations) for words or phrases in text content.

In this thesis, we use description logics (DLs) for symbolic representations, and in the following we present preliminaries of DLs and logic programming. We start to introduce syntax and semantics of DLs.

3.3.1 Description Logics

Description logics (DLs) [BCM$^+$03] correspond to a large fragment of standard ontology languages such as OWL. DLs can be used to represent knowledge of an application domain in a structured and formally well-understood way. In DLs, important notions of application areas are described by *names* for *concepts*, *roles*, and *attributes*. From a first-order logic point of view, concepts are unary, and roles as well as attributes are binary predicates to represent class membership conditions and arbitrary relations between two objects or object and a value from a concrete domain, respectively. Names will later be combined to complex description, and therefore names are also called atomic descriptions.

Let us assume that *Jumper*, *Event*, and *JumpingEvent* are selected as *atomic concept descriptions*, and *hasParticipant* is selected as an atomic role description by the knowledge modeler of the *athletics* domain. Assume further that the modeler would like to describe certain objects of the domain using these atomic descriptions, e.g., "An *Event* in which at least one *Jumper* participates." Then he needs concept *constructors* to build complex descriptions from atomic ones, e.g., $Event \sqcap (\exists_{\geq 1} hasParticipant.Jumper)$. It is also possible to *define* the atomic concept description *JumpingEvent* by stating that $JumpingEvent \doteq Event \sqcap (\exists_{\geq 1} hasParticipant.Jumper)$. We have a conjunction ($\sqcap$) and a constructor with $\exists$, which is called qualified cardinality restriction. Details will be explained below.

There are some variations of DLs suitable for different purposes. The prototypical description logic language $\mathcal{ALC}$ (Attributive Language with Complement) is the basis of many more expressive DLs, e.g., DLs with qualified cardinality restrictions ($\mathcal{Q}$). In this work, the DL $\mathcal{ALCQ}(\mathcal{D})$ with a concrete domain $\mathcal{D}$ for dealing with string values is introduced. We prefer qualified cardinality restrictions for modeling because atomic role descriptions can be selected that are more general, i.e., we do not need a role *hasJumperParticipant* in the example above.

Syntax

For a particular application a set C_N of atomic concepts (concept names) and a set R_N of atomic roles (role names) is assumed to be given. In $\mathcal{ALCQ}(\mathcal{D})$, descriptions for *complex concepts* are inductively built using concept constructors shown in Table 3.1, where $A \in C_N$ is a complex concept by definition, $R \in R_N$ is a role name, and $\mathcal{D}$ is a name for a decidable mathematical theory over a set of objects (e.g., linear inequations over real numbers or strings with equality as the only operator for comparing them). For concrete domains we assume another kind of descriptions be given, namely attribute names $\mathit{Attr} \in \mathit{Attr}_N$ and strings $s \in \Sigma^*$ for some alphabet Σ.

Syntax	Constructor
$C_1 \sqcap C_2$	conjunction
$C_1 \sqcup C_2$	disjunction
$\neg C$	negation
$\exists R.C$	existential restriction
$\forall R.C$	value restriction
$\exists_{\geq n} R.C$	qualified minimum restriction
$\exists_{\leq n} R.C$	qualified maximum restriction
$= \mathit{Attr}.s$	data type restriction
$\top$	top concept
$\bot$	bottom concept

Table 3.1: Constructors for building complex concepts in $\mathcal{ALCQ}(\mathcal{D})$.

For specifying semantic relations between complex concepts so called generalized concept inclusions (GCIs) of the form $C_1 \sqsubseteq C_2$ are used. A set of GCIs is called terminological box (T-box).

For instance with the GCI

$$\mathit{JumpingEvent} \sqsubseteq \mathit{Event} \sqcap \exists_{\geq 1} \mathit{hasParticipant}.\mathit{Jumper}$$

we specify that *JumpingEvent* is a specific event with the *necessary* condition that there is at least one related participant which is a jumper.

If we have both $A \sqsubseteq C$ and $C \sqsubseteq A$ in a *T-box* and A is not mentioned on the left-hand side of another GCI, we write $A \doteq C$ as an abbreviation. C is a necessary and *sufficient* condition for A in this case, hence we say A is *defined* and $A \doteq C$ is called a *concept definition*.

Specific objects considered in an application are referred to using individual names. So-called *assertions* are used to specify the following four cases:

Instance assertion: An individual name i is an *instance* of a concept name $A \in C_N$ is specified with the expression $i : A$.

Role assertion: Individual names i, j are in *relation* $R \in R_N$ is specified with $(i, j) : R$.

Attribute assertion: An individual name i has $s \in \Sigma^*$ as *value for attribute* $\mathit{Attr} \in \mathit{Attr}_N$ is specified with $(i, s) : \mathit{Attr}$.

Same-as assertion: Individual names i, j refer to the *same object* with $\mathit{same\text{-}as}(i, j)$.

A set of assertions is called an *A-box*. A pair $\mathcal{KB} = (\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is a T-box and $\mathcal{A}$ is an A-box is called a *knowledge base*. An *ontology* is a tuple $(C_N, R_N, \mathit{Attr}_N, \Sigma, \mathcal{T}, \mathcal{A})$.

Semantics

The semantics of a description logic knowledge base (and an ontology) is defined using an *interpretation* $\mathcal{I}$ that consist of a non-empty set Δ, the domain, and an interpretation function $\cdot^{\mathcal{I}}$, which assigns to every atomic concept description A a set $A^{\mathcal{I}} \subseteq \Delta$ and to every (atomic) role R a set $R^{\mathcal{I}} \subseteq \Delta \times \Delta$, every attribute Attr a set $\mathit{Attr}^{\mathcal{I}} \subseteq \Delta \times \Sigma^*$, $\top^{\mathcal{I}} = \Delta$, and $\bot^{\mathcal{I}} = \emptyset$. For complex concept descriptions, the interpretation function $\cdot^{\mathcal{I}}$ is extended as presented in Table 3.2.

The semantics of description logics is based on the notion of satisfiability.

An interpretation $\mathcal{I} = (\Delta, \cdot^{\mathcal{I}})$ *satisfies* a concept description C if $C^{\mathcal{I}} \neq \emptyset$. In this case, $\mathcal{I}$ is called a *model* for C. An interpretation $\mathcal{I}$ *satisfies* a GCI $C_1 \sqsubseteq C_2$ if $C_1^{\mathcal{I}} \subseteq C_2^{\mathcal{I}}$. An interpretation is a *model* of a T-box if it satisfies all GCIs in the T-box. A concept description C_1 *is subsumed by* a concept description C_2 w.r.t. a T-box if the GCI $C_1 \sqsubseteq C_2$ is satisfied in all models of

$$
\begin{array}{rcl}
(C_1 \sqcap C_2)^{\mathcal{I}} & = & C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}, \\
(C_1 \sqcup C_2)^{\mathcal{I}} & = & C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}, \\
(\neg C)^{\mathcal{I}} & = & \Delta \backslash C^{\mathcal{I}}\ , \\
(\exists R.C)^{\mathcal{I}} & = & \{x \in \Delta \mid \exists y \in \Delta \text{ with } (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\}, \\
(\forall R.C)^{\mathcal{I}} & = & \{x \in \Delta \mid \forall y \in \Delta, \text{ if } (x,y) \in R^{\mathcal{I}} \text{ then } y \in C^{\mathcal{I}}\}, \\
(\exists_{\geq n} R.C)^{\mathcal{I}} & = & \{x \in \Delta \mid \sharp\{y \in \Delta \mid (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\} \geq n\}, \\
(\exists_{\leq n} R.C)^{\mathcal{I}} & = & \{x \in \Delta \mid \sharp\{y \in \Delta \mid (x,y) \in R^{\mathcal{I}} \text{ and } y \in C^{\mathcal{I}}\} \leq n\}, \\
(= \mathit{Attr}.s)^{\mathcal{I}} & = & \{x \in \Delta \mid \exists (x,s) \in \mathit{Attr}^{\mathcal{I}}\} \\
 & & \text{where } \sharp M \text{ denotes the cardinality of the set } M.
\end{array}
$$

Table 3.2: Semantics of a description logic knowledge base.

the T-box. In this case, we also say that C_2 *subsumes* C_1. An interpretation $\mathcal{I}$ satisfies

- a concept assertion $i : A$ if $i^{\mathcal{I}} \in A^{\mathcal{I}}$,
- a role assertion $(i, j) : R$ if $(i^{\mathcal{I}}, j^{\mathcal{I}}) \in R^{\mathcal{I}}$,
- an attribute assertion $(i, s) : \mathit{Attr}$ if $(i^{\mathcal{I}}, s) \in \mathit{Attr}^{\mathcal{I}}$,
- a *same-as* assertion $\mathit{same\text{-}as}(i, j)$ if $i^{\mathcal{I}} = j^{\mathcal{I}}$.

A knowledge base ($\mathcal{T}$,$\mathcal{A}$) is *satisfied* by an interpretation $\mathcal{I}$ if $\mathcal{I}$ satisfies $\mathcal{T}$ and $\mathcal{A}$ (analogously for ontologies).

Let α be an assertion. A knowledge base $\mathcal{KB} = (\mathcal{T}, \mathcal{A})$ *entails* an assertion α (or: α follows from $\mathcal{KB}$), denoted as $\mathcal{KB} \models \alpha$ if for all models $\mathcal{I}$ of $\mathcal{KB}$ it holds that $\mathcal{I}$ satisfies α. Let $\mathcal{A}$ be an A-box. A knowledge base $\mathcal{KB}$ *entails* an A-box, denoted as $\mathcal{KB} \models \mathcal{A}$, if for all $\alpha \in \mathcal{A}$ it holds that $\mathcal{KB} \models \alpha$.

In the following sections we sometimes slightly misuse notation and assume that $(\mathcal{T}, \mathcal{A}) \cup \mathcal{A}'$ means $(\mathcal{T}, \mathcal{A} \cup \mathcal{A}')$. The function $\mathsf{inds}(\mathcal{A})$ delivers the individuals mentioned in A-box $\mathcal{A}$.

Decision problems and their reductions

A decision problem is a question with a true or false answer, depending on the values of some input parameters. The definitions given in the previous subsection can be paraphrased as decision problems:

- Check if a model for a concept description exists (*concept satisfiability* problem).
- Check if $C \sqsubseteq D$ holds in all models of a T-box (*concept subsumption* problem).
- Check if a model for a T-box exists (*T-box satisfiability* problem).

Satisfiability checks of content descriptions and consistency checks of A-boxes are useful to determine whether a knowledge base is meaningful at all. An overview about basic inference problems for A-boxes are given in the following:

- The *A-box consistency problem* for an A-box $\mathcal{A}$ (w.r.t. a T-box $\mathcal{T}$) is the problem of determining whether there exists a model of $\mathcal{A}$ (that is also a model of the T-box $\mathcal{T}$).
- Another problem is to test whether an individual i is an instance of a concept description C w.r.t. a T-box $\mathcal{T}$ and an A-box $\mathcal{A}$. A related problem is to test whether individuals i, j are related by role R w.r.t. a T-box $\mathcal{T}$ and an A-box $\mathcal{A}$ (*instance test* or *instance problem*: $(\mathcal{T}, \mathcal{A}) \models i : C$ or $(\mathcal{T}, \mathcal{A}) \models (i, j) : R$).
- The *instance retrieval* problem w.r.t. a query concept C and an ontology $\mathcal{O} = (C_N, R_N, Attr_N, \Sigma, \mathcal{T}, \mathcal{A})$ is to find all individuals i mentioned in the assertions of the A-box $\mathcal{A}$ such that i is an instance of C w.r.t. the T-box $\mathcal{T}$.

The following problem reductions are well known:

- The *concept satisfiability problem* for a concept description C can be reduced to the consistency problem for the A-box $\{i : C\}$.
- In order to solve the instance problem for an individual i and a concept description C one can check if the A-box $\{i : (\neg C)\}$ is inconsistent [BN03].
- In theory, the *retrieval problem* can be reduced to several instance problems.

In theory, all problems introduced above can be reduced to the A-box consistency problem. In practical systems, however, specific optimization techniques are used to decide a certain decision problem.

Queries for A-boxes $\mathcal{A}$

Assume that $A \in C_N, R \in R_N, Attr \in Attr_N$, $s \in \Sigma^*$, and V_N is a set of variable names (or variables for short), and let X, Y be variables or individual names, $A(X), R(X,Y), Attr(X,s), \textit{same-as}(X,Y)$ will be query atoms.

A substitution σ is mapping from $V_N \cup inds(\mathcal{A})$ to $inds(\mathcal{A})$ such that $\sigma(i) = i$ for $i \in inds(\mathcal{A})$.

A query is an expression $\{(X_1, \ldots, X_N) \mid \phi(X_1, \ldots, X_N)\}$ where $N \geq 0, \{X_1, \ldots, X_N\} \subseteq V_N$, and $\phi(\cdot)$ is a conjunction of atoms (the atoms are usually separated with commas) such that the variables in $\{X_1, \ldots, X_N\}$ are mentioned in at least one atom. The expression $(X_1, \ldots, X_N)$ is called query head, and $\phi(\cdot)$ is called query body (there can be additional variables in the body).

Given a query $cq = \{(X_1, \ldots, X_N) \mid \phi(X_1, \ldots, X_N)\}$, the query answering problem, called $answers((\mathcal{T}, \mathcal{A}), cq)$, is to compute all $(\sigma(X_1), \ldots, \sigma(X_n))$ such that for the atoms α in $\phi(\cdot)$ it holds that if $\alpha = A(X)$ then $(\mathcal{T}, \mathcal{A}) \models \sigma(X) : A$, if $\alpha = R(X,Y)$ then $(\mathcal{T}, \mathcal{A}) \models (\sigma(X), \sigma(Y)) : R$, if $\alpha = Attr(X,s)$ then $(\mathcal{T}, \mathcal{A}) \models (\sigma(X), s) : Attr$, and if $\alpha = \textit{same-as}(X,Y)$ then $(\mathcal{T}, \mathcal{A}) \models \textit{same-as}(\sigma(X), \sigma(Y))$, respectively. If for cq it holds that $N = 0$ then we have a so-called boolean query. We say the query is answered with *true* if $answers((\mathcal{T}, \mathcal{A}), cq))= \{()\}$ and *false* otherwise ($answers((\mathcal{T}, \mathcal{A}), cq)=\{\}$).

Before we describe the syntax and semantics of another modeling construct, namely a rule, we need preliminaries of logic programming.

3.3.2 Logic Programming

Logic Programming uses the language of logic to express data and programs - in most cases first-order logic (FOL). Similarly to first-order logic, logic programming allows for constant, function and predicate symbols. *Atomic formulas* (also known as *atomic sentences* or *atoms* in short) have the form $p(t_1, \ldots, t_n)$, where the t_i are terms and p is a predicate symbol of arity n. An atomic formula or its negation is called a *literal.* A *clause* is a logic formula of the form

$$L_1 \vee \ldots \vee L_n,\ n \geq 0 \tag{3.16}$$

where each L_i is a literal. A *Horn clause* is a clause that contains at most one positive literal. A *definite clause* (also known as *rule*) is a Horn clause that contains exactly one positive literal. A Horn clause without negative literals is called a *fact.*

Following the notational convention proposed e.g. in [NM95], definite clauses are written as follows:

$$A_0 \leftarrow A_1, \ldots, A_n. \tag{3.17}$$

where $n \geq 0$ and $A_0, \ldots, A_n$ are atomic formulas. All variables occurring in a formula are universally quantified over the whole formula. The backward arrow ← is read as "if", and "," as "and." The atomic formula A_0 is called the *head* of the clause whereas the sequence of $A_1, \ldots, A_n$ is called the *body* of the clause. If $n = 0$, then the body is equivalent to *true*, and the clause $A_0 \leftarrow true$ is abbreviated to A_0 and is called a *fact.* Otherwise if $n \neq 0$, the clause is called a *rule* [Kow88]. Formulas and clauses are called *ground* if they contain no variables. A fact is a ground atomic formula. A Horn clause with an empty head, i.e. where A_0 is absent, is called a *goal clause.* A *definite program* is a finite set of rules or facts. A program is *recursive* if the body of one rule directly or indirectly depends on the head of another rule, otherwise it is called *non-recursive*.

To give an example, let Π be a definite program containing given with the following rules and facts:

$$\begin{aligned}
&q(x) \leftarrow p(x).\\
&r(x) \leftarrow q(x).\\
&p(i).\\
&r(j).
\end{aligned}$$

A definite program with variables can be considered as a shorthand for the set of all ground instances of its rules, i.e., for the result of substituting variables in the rules of the program in all possible ways (this process is often referred to as grounding). Therefore Π is a shorthand for the following set of ground instances of its rules plus the above mentioned facts:

$$\begin{aligned} q(i) &\leftarrow p(i). \\ r(i) &\leftarrow q(i). \\ q(j) &\leftarrow p(j). \\ r(j) &\leftarrow q(j). \end{aligned}$$

The set of all ground atoms in the language of a definite program Π is called the Herbrand base of Π and denoted by *HB*. Note that the language of a definite program is the set of constant, function and predicate symbols that occur in the definite program. In our example Π has the following Herbrand base

$$HB = \{p(i), q(i), r(i), p(j), q(j), r(j)\}. \tag{3.18}$$

A *Herbrand model* of a definite program Π is a subset of *HB*. The semantics of a definite program Π is the smallest Herbrand model.

Datalog, a prominent query and rule language used in deductive databases, supports only definite clauses without function symbols. In addition, Datalog requires all variables that appear in the head of a rule to appear also in the body of the same rule[3]. Systems supporting Datalog often employ *forward-chaining*, also known as bottom-up inference. Here the name forward-chaining indicates that rules are processed forward, i.e., in the sense of the logical implication sign, from body (premise) to head (conclusion).

Prolog is a widely-used logic programming system, and, unlike Datalog, Prolog supports definite clauses with function symbols. Prolog uses resolution based inference algorithms, which work in a *backward-chaining* way, also known as top-down or goal-directed inference. Backward-chaining inference based on the SLD resolution does not always guarantee termination since inference with definite clauses with function symbols is undecidable in general [VEK76], but decidable for certain subclasses, whereas termination is guaranteed for the fixed-point based inference algorithms employed for Datalog [Ull85].

[3] safety property

In the context of Datalog, one usually distinguishes between two sets of clauses: a set of ground facts, called the Extensional Database (EDB), and a set of rules or Datalog program Π, called the Intentional Database (IDB). The predicates that appear in the EDB are called EDB-predicates. EDB-predicates may appear in Π as well, but only in clause bodies. The predicates that appear in Π but not in the EDB are called IDB-predicates . As a consequence, the head predicate of each clause in Π is an IDB-predicate.

Assume that a Datalog rule $A_0 \leftarrow A_1, \ldots, A_n$ and a set of ground facts $F =$ EDB are given. If a substitution θ exists, which replaces variables with constants, such that for each $1 \leq i \leq n$ it holds that $\theta(A_i) \in F$, i.e. the premises of the rule are satisfied, then we can infer the fact $\theta(A_0)$, also known as the conclusion. In other words, we say that a rule is applied to a set of ground facts or extensional knowledge base (EDB). Notice that the inferred fact may either be a new fact or it may already be contained in the EDB. As mentioned above, we say a set of rules Π is applied to a KB in a forward-chaining way, if for every rule in Π whose premises are satisfied the conclusion of the rule is added to the EDB, and this process is repeated until a fixed point is reached such that no new facts can be added to the EDB. As an example, consider the following set of ground facts stored as tuples in a relational database:

$$EDB = \{parent(mary, john), parent(john, michael)\}$$

and the following Datalog program Π consisting of the two rules:

$$\begin{aligned} ancestor(i, j) &\leftarrow parent(i, j) \\ ancestor(i, j) &\leftarrow parent(i, k), ancestor(k, j) \end{aligned}$$

that defines the ancestor relationship. As a consequence of applying the rules in Π to EDB, the following facts are added to the EDB:

$$\{ancestor(mary, john), ancestor(john, michael), ancestor(mary, michael)\}.$$

Notice that after the addition of these tuples to the EDB, a fixed point is reached and no new facts can be inferred.

In this thesis we use Datalog rules in a backward-chaining way to define a space of possible latent structures used as symbolic descriptions (annotations) of media content. As an example we assume that the following rule is given:

$$\begin{aligned} isAdjacent(Y,Z) \leftarrow & PoleVault(X), hasPart(X,Y), Crossbar(Y), \\ & hasPart(X,W), Pole(W), hasParticipant(X,Z) \\ & PoleVaulter(Z). \end{aligned}$$

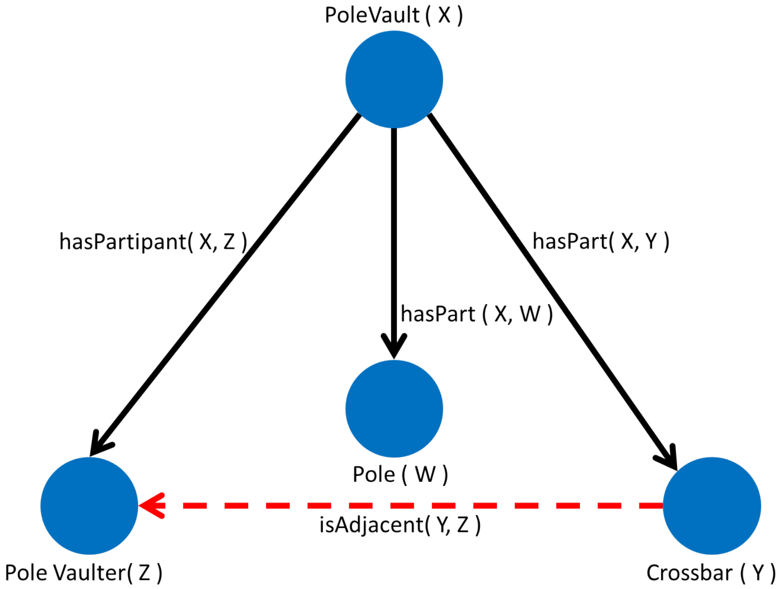


Figure 3.2: **Example for a triangular structure.** For an atom $isAdjacent(i,j)$ new objects are generated, namely a pole vault X and a pole W, with respective relations to the pole vaulter $Z = j$ and the crossbar $Y = i$.

The idea is that for an atom $isAdjacent(i,j)$ the rule given above can be employed in a backward-chaining way, and, as shown in Figure 3.2, possibly new objects will be generated, namely a pole vault X and a pole W, with respective relations to the pole vaulter $Z = j$ and the crossbar $Y = i$.

With backward-chaining for atoms α, new atoms are generated such that α atoms can be derived in a forward-chaining way if the new atoms were added to the EDB. The newly generated atoms serve as explanations for α atoms, and backward-chaining in this way can be seen as a form of abduction.

Computing explanations via abduction

Abduction [Sha05, PKMM08, HB12, MGXC12, Sob13] can be considered as a new type of non-standard inference service. In this view, observations (or parts of them) are utilized to constitute A-box entailment problems. More formally, for a given set of A-box assertions Γ (observations, to be explained) and a knowledge base $\mathcal{KB} = (\mathcal{T}, \mathcal{A})$, the abductive retrieval inference service aims to derive all sets of A-box assertions Δ (explanations) such that

$$\mathcal{KB} \cup \Delta \models \Gamma \tag{3.19}$$

and the following conditions are satisfied:

- $\Delta \in \mathit{backward\_chain}(\mathcal{T}, \mathcal{A}, \mathcal{R}, \Gamma) \wedge S(\Delta) > 0$, where S is a monotone scoring function.
- $\mathcal{KB} \cup \Delta$ is satisfiable, and
- Δ is a minimal explanation for Γ, i.e., there exists no other explanation $\Delta' \subsetneq \Delta$ and it holds that $\mathcal{KB} \cup \Delta' \models \Gamma$.

The function *backward_chain* is defined as follows:

$$\mathit{backward\_chain}(\mathcal{T}, \mathcal{A}, \mathcal{R}, \Gamma) = \bigcup_{\gamma \in \Gamma \wedge \mathit{requires\_fiat}(\gamma)} \mathit{bc}(\mathcal{T}, \mathcal{A}, \mathcal{R}, \gamma) \tag{3.20}$$

In turn, *bc* is defined as

$$\mathit{bc}(\mathcal{T}, \mathcal{A}, \mathcal{R}, (Z) : P) = \mathit{transform}(\Phi, \sigma) \tag{3.21}$$

if there exists a rule

$$r = P(X) \leftarrow Q_1(Y_1), \ldots, Q_n(Y_n) \in \mathcal{R} \tag{3.22}$$

such that a set of query atoms Φ and an admissible variable substitution σ with $\sigma(X) = Z$ can be found, and the query

$$\{() \mid \mathit{expand}(P(Z), r, \mathcal{R}) \setminus \Phi\} \tag{3.23}$$

is answered with *true*. Otherwise, or if no such rule r exists in $\mathcal{R}$, it holds that

$$\mathit{bc}(\mathcal{T}, \mathcal{A}, \mathcal{R}, (Z) : P) = \{(Z) : P\}. \tag{3.24}$$

The goal of the function *backward_chain* is to determine what must be added (Φ) such that an entailment

$$\mathcal{KB} \cup \Gamma \cup \Phi \models (Z) : P \tag{3.25}$$

holds. The set of query atoms Φ defines what must be hypothesized in order to answer the query Q with *true* such that

$$\Phi \subseteq \mathit{expand}(P(Z), r, \mathcal{R}) \tag{3.26}$$

holds. The definition of *backward_chain* is non-deterministic due to several possible choices for Φ. The function application

$$\mathit{expand}(P(Z), P(X) \leftarrow Q_1(Y_1), \ldots, Q_n(Y_n), \mathcal{R}) \tag{3.27}$$

is also defined in a non-deterministic way as

$$\mathit{expand}'(\sigma'(Q_1(Y_1)), \mathcal{R}) \cup \cdots \cup \mathit{expand}'(\sigma'(Q_n(Y_n)), \mathcal{R}) \tag{3.28}$$

where σ' is a minimal substitution such that $\sigma'(X) = Z$ and $\mathit{expand}'(P(X), \mathcal{R})$ being $\mathit{expand}(P(X), r, \mathcal{R})$ if there exist a rule $r = P(X) \leftarrow \ldots \in \mathcal{R}$ and $\langle P(X) \rangle$ otherwise. We say the set of rules is backward-chained, and since there might be multiple rules in $\mathcal{R}$, backward chaining is non-deterministic.

The function *requires_fiat* depends on the application context and we assume the following definition:

$$\mathit{requires\_fiat}((Z) : P) = \mathit{true} \text{ iff } P \in \{\mathit{near}, \mathit{adjacent\_to}, \ldots\} \tag{3.29}$$

Obviously, backward chaining potentially produces infinite structures if rules are recursive. Therefore we assume a scoring function S to be applied to all Δs, with the goal to inhibit infinite structures. For implementing A-box interpretations we use a function $\mathit{interpret}(\mathcal{T}, \mathcal{A}, \mathcal{R}, S, \Gamma)$ defined as

$$\mathit{maximize}(\{\Delta \mid \Delta \; \mathit{is\ an\ explanation}\}, S), \tag{3.30}$$

where S is a scoring function defined as follows

$$S(\Delta) := S_i(\Delta) - S_h(\Delta), \tag{3.31}$$

where S_i and S_h are defined as follows [PKMM08]:

$$S_i := |\{i | i \in \mathit{inds}(\Delta) \text{ and } i \in \mathit{inds}(\Phi)\}| \tag{3.32}$$

$$S_h := |\{i | i \in \mathit{inds}(\Delta) \text{ and } i \in \mathit{newInds}\}| \tag{3.33}$$

The set *newInds* contains all individuals that are hypothesized during the generation of an explanation (new individuals). An explanation with the highest score is preferred to others. Kaya has presented in [Kay11] that a scoring function is required because of possibly existing recursive rules, but the A-box abduction algorithm does not provide a depth or branch control for sequential abduction steps. In order to avoid the limitations of A-box abduction and the scoring function, in [Naf13] Nafissi suggest to use a probabilistic scoring function for interpretation purposes. The new function is implemented as an extension to [Kay11]. In this thesis we use Kaya's scoring function for explaining the interpretation process as a part of the symbolic content description approach because the observations are strict, and we consider only a single abduction step so that we have no problems with recursive rules.

In the context of symbolic IR, A-boxes are symbolic representations for document modalities such as images, text, and caption. It may be possible that assertions in an A-box do not precisely enough represent the content of document modalities. A-box abduction is one technique which can be used for creating interpretations (new assertions) so that symbolic IR systems can deliver documents with higher precision.

A-box difference

In general, one document consists of different modalities. Each modality has at least one symbolic representation. The fusion of all representations for a document can lead to higher precision results [Kay11]. In this thesis we present an implementation of fusion using the so-called *A-box difference operator* in Section 4.2. The A-box difference operator defined in [HMW07, MOH$^+$16] provides means for determining differences on a semantic basis.

The *semantic difference* $\Delta_{\mathcal{A},\mathcal{B}}$ (read as: Δ from $\mathcal{A}$ added to $\mathcal{B}$ so that $\mathcal{A}$ is entailed) of two A-boxes $\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$ and $\mathcal{B} = \{\beta_1, \ldots, \beta_m\}$ is a set of assertions of $\mathcal{A}$ with such that:

1. There exists a (not necessarily total) mapping ϕ : $\mathsf{inds}(\mathcal{B}) \mapsto \mathsf{inds}(\mathcal{A})$ such that $(\mathcal{T}, \phi(\mathcal{B}) \cup \Delta_{\mathcal{A},\mathcal{B}}) \models \mathcal{A}$ where ϕ is defined as follows: $\phi(\mathcal{B}) =_{def} \{\phi(\beta_1), \ldots, \phi(\beta_m)\}$, $\phi(i : C) =_{def} \phi(i) : C$, $\phi((i, j) : R) =_{def} (\phi(i), \phi(j)) : R$. (Analogously for attribute and same-as assertions.)

2. $\phi(\mathcal{B}) \cup \Delta_{\mathcal{A},\mathcal{B}}$ is satisfiable w.r.t. $\mathcal{T}$,

3. $\Delta_{\mathcal{A},\mathcal{B}} \subseteq \mathcal{A}$ is minimal.

In the following we use a function term $\mathit{abox\_diff}(\mathcal{T}, \mathcal{A}, \mathcal{B})$ to denote $(\Delta_{\mathcal{A},\mathcal{B}}, \phi(\mathcal{B}))$. Note that the A-box difference operator is not commutative.

Example 3.1 *The T-box* $\mathcal{T} = \{Person \doteq \exists hasPart.Body \sqcap \exists hasPart.Face\}$*, the A-box* $\mathcal{A} = \{i : (\exists hasPart.Body \sqcap \exists hasPart.Face)\}$*, and the A-box* $\mathcal{B} = \{j : Person\}$ *are given.*

The A-box differences are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{\}$ because $\phi(\mathcal{B}) = \{i \mapsto j\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{i : (\exists hasPart.Body \sqcap \exists\ hasPart.Face)\})) \models \{j : Person\}$.

- $\Delta_{\mathcal{B},\mathcal{A}} = \{j : Person\}$ because $\phi(\mathcal{A}) = \{\}$, so that a new assertion is hypothesized.

Example 3.2 *The T-box* $\mathcal{T} = \{\}$*, the A-box* $\mathcal{A} = \{i : Athlete\}$*, and the A-box* $\mathcal{B} = \{j : Athlete\}$ *are given.*

The A-box differences are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{\}$ because $\phi(\mathcal{B}) = \{i \mapsto j\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{j : Athlete\})) \models \{i : Athlete\}$.

- $\Delta_{\mathcal{B},\mathcal{A}} = \{\}$ because $\phi(\mathcal{A}) = \{j \mapsto i\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{i : Athlete\})) \models \{j : Athlete\}$.

Example 3.3 *The T-box* $\mathcal{T} = \{\}$*, the A-box* $\mathcal{A} = \{i : Athlete\}$*, and the A-box* $\mathcal{B} = \{j : Person\}$ *are given.*

The A-box differences are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{i : \mathit{Athlete}\}$ because $\phi(\mathcal{B}) = \{\}$, so that a new assertion is hypothesized.

- $\Delta_{\mathcal{B},\mathcal{A}} = \{j : \mathit{Person}\}$ because $\phi(\mathcal{A}) = \{\}$, so that a new assertion is hypothesized.

Previous examples present that the A-box difference operator fulfills the first condition. The next example presents that the first condition is fulfilled but the second condition $\phi(\mathcal{B}) \cup \Delta_{\mathcal{A},\mathcal{B}}$ w.r.t. $\mathcal{T}$ is not satisfiable.

Example 3.4 *The T-box $\mathcal{T} = \{\}$, the A-box $\mathcal{A} = \{i : \neg \mathit{Athlete}\}$, and the A-box $\mathcal{B} = \{j : \mathit{Athlete}\}$ are given.*

The returned results of the A-box difference operator are:

- $(\Delta_{\mathcal{A},\mathcal{B}}, \phi(\mathcal{B})) = \{\}$ because $\phi(\beta_1) = \{k \mapsto i\}$ with $\Delta_{\mathcal{A},\mathcal{B}} = \{k : \neg \mathit{Athlete}\}$, since $\phi(\beta_2) = \{j \mapsto i\}$ with $\Delta_{\mathcal{A},\mathcal{B}} = \{j : \neg \mathit{Athlete}\}$ violates condition 2, since $\{j : \neg \mathit{Athlete}\} \cup \{j : \mathit{Athlete}\}$ is inconsistent.

- $\Delta_{\mathcal{B},\mathcal{A}} = \{j : \neg \mathit{Athlete}\}$ because $\phi(\mathcal{A}) = \{\}$, so that a new assertion is hypothesized.

In the following further examples are presented with varying assertions.

Example 3.5 *The T-box $\mathcal{T} = \{\mathit{Person} \doteq \exists \mathit{hasPart}.\mathit{Face} \sqcap \exists \mathit{hasPart}.\mathit{Body}\}$, the A-box $\mathcal{A} = \{i : \mathit{Person}\}$, and the A-box $\mathcal{B} = \{j : \exists \mathit{hasPart}.\mathit{Face}\}$ are given.*

The A-box differences are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{i : \mathit{Person}\}$ because $\phi(\mathcal{B}) = \{\}$, so that a new assertion is hypothesized.

- $\Delta_{\mathcal{B},\mathcal{A}} = \{\}$ because $\phi(\mathcal{A}) = \{j \mapsto i\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{i : \mathit{Face}\})) \models \{j : \mathit{Person}\}$.

Example 3.6 *The T-box $\mathcal{T} = \{\mathit{Person} \doteq \exists \mathit{hasPart}.\mathit{Face} \sqcap \exists \mathit{hasPart}.\mathit{Body}\}$, the A-box $\mathcal{A} = \{i : \mathit{Person}\}$, and the A-box $\mathcal{B} = \{j : \exists \mathit{hasPart}.\mathit{Face}, k : \exists \mathit{hasPart}.\mathit{Body}\}$ are given.*

The results are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{i : Person\}$ $\phi(\mathcal{B}) = \{\}$, so that a new assertion is hypothesized.
- $\Delta_{\mathcal{B},\mathcal{A}} = \{j : \exists hasPart.Face, k : \exists hasPart.Body\}$ $\phi(\mathcal{A}) = \{\}$, so that new assertions are hypothesized.

Example 3.7 *The T-box $\mathcal{T} = \{Athlete \doteqdot Person \sqcap participatesIn.SportsEvent\}$, the A-box $\mathcal{A} = \{i : Athlete\}$, and the A-box $\mathcal{B} = \{j : Person\}$ are given.*

The results are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{j : Athlete\}$, because $\phi(\mathcal{B}) = \{\}$.
- $\Delta_{\mathcal{B},\mathcal{A}} = \{\}$ because $\phi(\mathcal{A}) = \{j \mapsto i\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{i : Person\})) \models \{j : Athlete\}$.

Example 3.8 *$\mathcal{T} = \{Athlete \doteqdot Person \sqcap participatesIn.SportsEvent\}$, $\mathcal{A} = \{i : Athlete\}$, and $\mathcal{B} = \{j : Person, (j,k) : participatesIn, k : SportsEvent\}$ are given.*

The results are:

- $\Delta_{\mathcal{A},\mathcal{B}} = \{i : Athlete\}$ because for the mappings $\phi(\beta_1) = \{k \mapsto i\}$ and $\phi(\beta_2) = \{j \mapsto i\}$ there are not entailed assertions.
- $\Delta_{\mathcal{B},\mathcal{A}} = \{(j,k) : participatesIn, k : SportsEvent\}$ which has two hypothesized assertions.

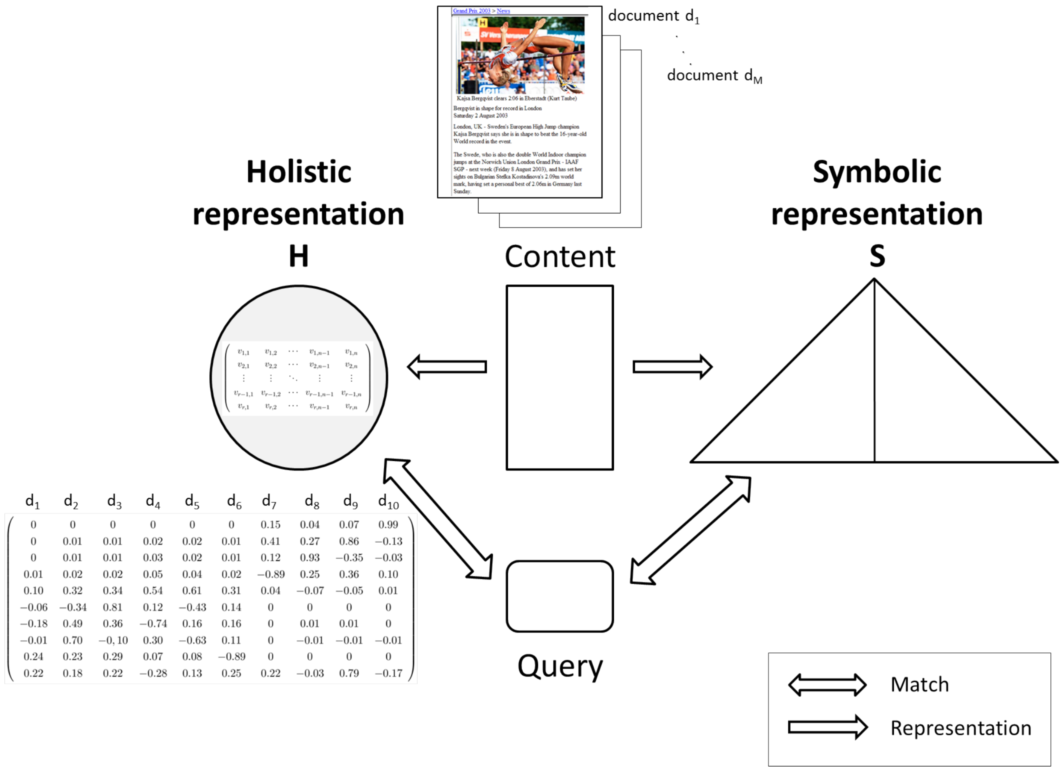
Chapter 4

Holistic and Symbolic Content Retrieval

In the previous chapter we have presented fundamentals for holistic and symbolic representations in order to formalize a context-specific environment to be used in Nonaka's knowledge creation process. In this chapter we describe holistic and symbolic content descriptions as well as retrieval processes using a representative example for our knowledge management scenario with the aim of presenting precision and recall results for both approaches, before we present our methodology for systematically combining holistic and symbolic content descriptions deriving a formalization for the knowledge creation process (Chapter 5).

4.1 Holistic Content Description Approach

A holistic content description approach, here the LSI approach, is used in order to create holistic content descriptions (see Figure 4.1 [left]). As explained above LSI includes singular value decomposition (SVD) and latent semantic indexing as a technique for better support information retrieval. It is also possible to use other holistic approaches, such as, e.g., LDA as discussed in Section 3.2.2. However, the key point for our choice is that holistic information retrieval supports high recall and provides a means to compute latent structures.



Grand Prix 2003 > News

Kajsa Bergqvist clears 2:06 in Eberstadt (Kurt Taube)

Bergqvist in shape for record in London
Saturday 2 August 2003

London, UK - Sweden's European High Jump champion Kajsa Bergqvist says she is in shape to beat the 16-year-old World record in the event.

The Swede, who is also the double World Indoor champion jumps at the Norwich Union London Grand Prix - IAAF SGP - next week (Friday 8 August 2003), and has set her sights on Bulgarian Stefka Kostadinova's 2.09m world mark, having set a personal best of 2.06m in Germany last Sunday.

$$\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n-1} & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{r-1,1} & v_{r-1,2} & \cdots & v_{r-1,n-1} & v_{r-1,n} \\ v_{r,1} & v_{r,2} & \cdots & v_{r,n-1} & v_{r,n} \end{pmatrix}$$

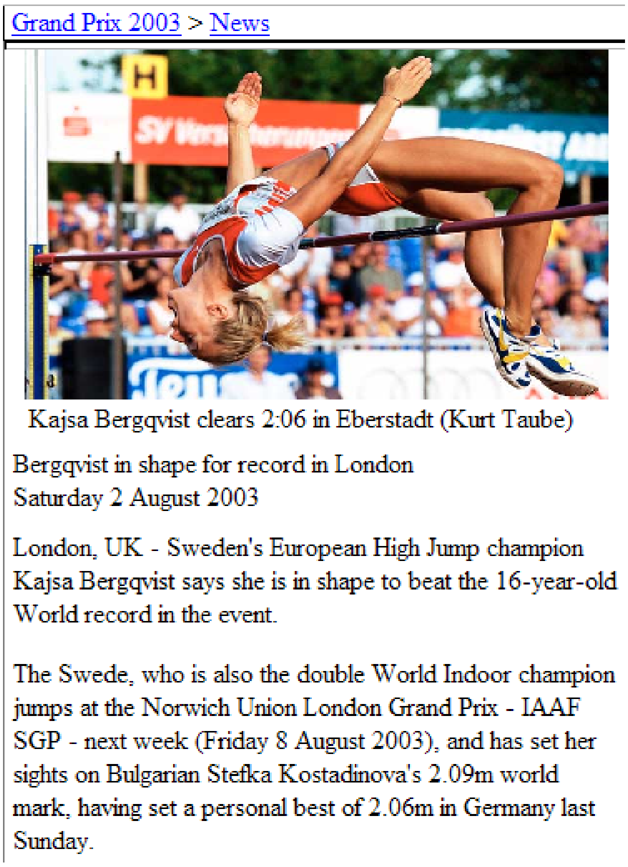
$$\begin{array}{c} \begin{array}{cccccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.04 & 0.07 & 0.99 \\ 0 & 0.01 & 0.01 & 0.02 & 0.02 & 0.01 & 0.41 & 0.27 & 0.86 & -0.13 \\ 0 & 0.01 & 0.01 & 0.03 & 0.02 & 0.01 & 0.12 & 0.93 & -0.35 & -0.03 \\ 0.01 & 0.02 & 0.02 & 0.05 & 0.04 & 0.02 & -0.89 & 0.25 & 0.36 & 0.10 \\ 0.10 & 0.32 & 0.34 & 0.54 & 0.61 & 0.31 & 0.04 & -0.07 & -0.05 & 0.01 \\ -0.06 & -0.34 & 0.81 & 0.12 & -0.43 & 0.14 & 0 & 0 & 0 & 0 \\ -0.18 & 0.49 & 0.36 & -0.74 & 0.16 & 0.16 & 0 & 0.01 & 0.01 & 0 \\ -0.01 & 0.70 & -0,10 & 0.30 & -0.63 & 0.11 & 0 & -0.01 & -0.01 & -0.01 \\ 0.24 & 0.23 & 0.29 & 0.07 & 0.08 & -0.89 & 0 & 0 & 0 & 0 \\ 0.22 & 0.18 & 0.22 & -0.28 & 0.13 & 0.25 & 0.22 & -0.03 & 0.79 & -0.17 \end{pmatrix} \end{array}$$

Figure 4.1: **Holistic representation of document contents**. The columns of the matrix H are the holistic representation for the documents $d_1 \dots d_{10}$.

4.1.1 Holistic Content Description Creation Process

For explaining how holistic content descriptions are created, we return to the knowledge management scenario presented in Section 2.2.2. An engineer has the task to construct athletic areas, for instance, jumping areas in a stadium. As we have argued in Section 2.2.2, standard retrieval techniques do deliver context-specific results neither automatically nor in a proactive way because of missing latent structures for a document. How to compute the latent structure of documents in a repository is described in the following using LSI.

Assume that we have a repository which contain documents from different domains such as an *Athletics* domain and a *FairyTale* domain. In particular, assume that we have ten documents: Six documents are from the *Athletics* domain (labeled $d_1 \dots d_6$) and four documents are from the *FairyTale* domain (labeled $d_7 \dots d_{10}$). Each document contains images, caption texts, and texts in general. Document d_1 is shown in Figure 4.2. In the text we can find that d_1 describes high jump events. For the holistic approach we only use the textual parts (caption and text) of a document. Titles of the sample dataset of documents are presented in Table 4.1.



Grand Prix 2003 > News

Kajsa Bergqvist clears 2:06 in Eberstadt (Kurt Taube)

Bergqvist in shape for record in London
Saturday 2 August 2003

London, UK - Sweden's European High Jump champion Kajsa Bergqvist says she is in shape to beat the 16-year-old World record in the event.

The Swede, who is also the double World Indoor champion jumps at the Norwich Union London Grand Prix - IAAF SGP - next week (Friday 8 August 2003), and has set her sights on Bulgarian Stefka Kostadinova's 2.09m world mark, having set a personal best of 2.06m in Germany last Sunday.

Figure 4.2: **A sample multimedia document d_1 with athletics news.** Source: IAAF [IAA09]

For demonstration purposes we use the thirteen words *high*, *jump*, *long*, *pole*, *vault*, *person*, *snow*, *white*, *prince*, *charming*, *sleeping*, *beauty*, and *Rapunzel* for indexing as representative terms for our knowledge management scenario. The words *high*, *jump*, *long*, *pole*, *vault*, and *person* are often used in articles for describing sports events. The words which are often used to describe fairy tales are *snow*, *white*, *prince*, *charming*, *sleeping*, *beauty*, and *Rapunzel*. How to automatically find topics for describing athletics news or fairy tales is described in [Ble12].

For presenting the LSI approach, in our example the input term-document matrix is a 13 × 10 matrix C for the ten documents $d_1 \dots d_{10}$ (see Table 4.1) and

Document number	Title	Origin
d_1	Kajsa Bergqvist clears 2:06 in Eberstadt	London 2003 2 August 2003 [IAA09]
d_2	Women Pole Vault Qualification	Helsinki 2005 News Team, 7 August 2005 [IAA09]
d_3	Silnov improves to 2.37	Bob Ramsak for the IAAF, 20 August 2006 [IAA09]
d_4	Lysenki closes in on World record with 75.95m Hammer Throw	(c) 1996-2007 IAAF, 14 July 2005 [IAA09]
d_5	Mack delights with 6.01 vault	Bob Ramsak for the IAAF 18 September 2004 [IAA09]
d_6	Gay skims 200m in 19.79	Chris Turner for the IAAF, 25 August 2006 [IAA09]
d_7	Prince Charming	[Wik15b]
d_8	Rapunzel	[Wik15c]
d_9	Sleeping Beauty	[Wik15d]
d_{10}	Snow White	[Wik15e]

Table 4.1: **Representative documents from the *Athletics* and *FairyTale* domain.**

the thirteen terms mentioned above. The entries in the term-document matrix are simply the frequencies (counts) with which a term occurs in the respective document. In Table 4.2 is presented that the documents d_1, d_3, d_4, d_5, and d_6 about athletics events have no words which are used for describing fairy tales. Document d_2 only contains the word "white." Words such as *long* and *jump* for describing athletics news can be found in *Athletics* documents ($d_1 \ldots d_6$) as well as in *FairyTale* documents ($d_7 \ldots d_{10}$). But in both document categories words are differently combined. In athletics documents the term "long" is often used with the term "jump" whereas in fairy tales "for a long time" is a coherent term. The idea of LSI is to compute a latent structure such that for the string queries "long jump" and "high jump" a user receives no documents about fairy tales and for the string query "snow white" he obtains no documents about athletics. The representation of each document w.r.t. a given repository is given by the columns of a document matrix V^T. For the input term-document matrix the

Term/document	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
high	1	2	2	5	7	3	0	0	1	0
jump	2	5	4	10	9	4	4	15	19	1
long	0	1	3	3	1	3	1	7	5	1
pole	0	3	1	0	2	1	0	0	0	0
vault	0	4	1	0	5	2	0	0	0	0
person	0	0	7	2	1	2	1	0	1	0
snow	0	0	0	0	0	0	13	1	1	123
white	0	1	0	0	0	0	13	2	1	123
prince	0	0	0	0	0	0	48	21	48	13
charming	0	0	0	0	0	0	37	2	0	1
sleeping	0	0	0	0	0	0	7	1	56	2
beauty	0	0	0	0	0	0	7	2	52	2
Rapunzel	0	0	0	0	0	0	1	58	0	0

Table 4.2: **Term-document matrix for a corpus with texts from the *Athletics* ($d_1 \dots d_6$) and** *emphFairyTale* **domains ($d_7 \dots d_{10}$).** Representative terms are *high*, *jump*, *long*, *pole*, *vault*, *person*, *snow*, *white*, *prince*, *charming*, *sleeping*, *beauty*, and *Rapunzel*. Each matrix cell indicates the frequency with which a term occurs in the respective document.

according matrices are a term matrix U, a matrix Σ with the singular values, and a document matrix V^T. For our examples the three matrices are presented in the following:

$$U = \begin{pmatrix} 0 & 0.01 & 0 & 0.02 & 0.48 & -0.16 & -0.13 & -0.65 & -0.32 & 0,26 \\ 0.02 & 0.22 & 0.14 & 0.18 & 0.71 & -0.11 & -0.35 & 0.31 & 0.35 & 0.21 \\ 0.01 & 0.07 & 0.08 & 0.07 & 0.2 & 0.35 & 0 & 0.43 & -0.8 & -0.21 \\ 0 & 0 & 0 & 0 & 0.15 & -0.14 & 0.48 & 0.38 & 0.15 & 0.41 \\ 0 & 0 & 0 & 0.01 & 0.27 & -0.35 & 0.72 & -0.12 & -0.12 & -0.1 \\ 0 & 0.01 & 0 & 0 & 0.24 & 0.83 & 0.31 & -0.24 & 0.3 & -0.1 \\ 0.70 & -0.10 & -0.02 & 0.03 & -0.01 & 0.02 & -0.05 & -0.16 & -0.07 & -0.03 \\ 0.70 & -0.10 & 0 & 0.04 & 0.01 & -0.03 & 0.05 & 0.16 & 0.07 & -0.38 \\ 0.14 & 0.64 & 0.13 & -0.42 & -0.07 & -0.01 & 0.03 & -0.05 & -0.02 & 0.39 \\ 0.04 & 0.15 & 0.10 & -0.72 & 0.08 & -0.01 & -0.03 & 0.04 & 0.00 & 0.05 \\ 0.04 & 0.50 & -0.30 & 0.31 & -0.12 & 0.00 & 0.05 & -0.05 & -0.02 & 0.55 \\ 0.04 & 0.47 & -0.26 & 0.29 & -0.11 & 0.00 & 0.05 & -0.05 & -0.02 & 0.22 \\ 0.01 & 0.16 & 0.89 & 0.30 & -0.20 & -0.01 & 0.08 & -0.12 & 0.01 & -0.12 \end{pmatrix}.$$

$$\Sigma = \begin{pmatrix} 176.47 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 101.72 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 60.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 45.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 19.35 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6.92 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.78 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.19 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.62 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0001 \end{pmatrix}.$$

$$V^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.04 & 0.07 & 0.99 \\ 0 & 0.01 & 0.01 & 0.02 & 0.02 & 0.01 & 0.41 & 0.27 & 0.86 & -0.13 \\ 0 & 0.01 & 0.01 & 0.03 & 0.02 & 0.01 & 0.12 & 0.93 & -0.35 & -0.03 \\ 0.01 & 0.02 & 0.02 & 0.05 & 0.04 & 0.02 & -0.89 & 0.25 & 0.36 & 0.10 \\ 0.10 & 0.32 & 0.34 & 0.54 & 0.61 & 0.31 & 0.04 & -0.07 & -0.05 & 0.01 \\ -0.06 & -0.34 & 0.81 & 0.12 & -0.43 & 0.14 & 0 & 0 & 0 & 0 \\ -0.18 & 0.49 & 0.36 & -0.74 & 0.16 & 0.16 & 0 & 0.01 & 0.01 & 0 \\ -0.01 & 0.70 & -0,10 & 0.30 & -0.63 & 0.11 & 0 & -0.01 & -0.01 & -0.01 \\ 0.24 & 0.23 & 0.29 & 0.07 & 0.08 & -0.89 & 0 & 0 & 0 & 0 \\ 0.22 & 0.18 & 0.22 & -0.28 & 0.13 & 0.25 & 0.22 & -0.03 & 0.79 & -0.17 \end{pmatrix}$$

Computing an approximation for C by keeping the first two singular values and using the corresponding columns from the U and V matrices yields $C \approx C_2 = U_2 \Sigma_2 V_2^T$:

$$U_2 = \begin{pmatrix} 0 & 0.01 \\ 0.02 & 0.22 \\ 0.01 & 0.07 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0.01 \\ 0.70 & -0.10 \\ 0.70 & -0.10 \\ 0.14 & 0.64 \\ 0.04 & 0.15 \\ 0.04 & 0.50 \\ 0.04 & 0.47 \\ 0.01 & 0.16 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 176.47 & 0 \\ 0 & 101.72 \end{pmatrix}$$

$$V_2^T = \begin{pmatrix} 0 & 0 & 0 & 0.01 & 0.10 & -0.06 & -0.18 & -0.01 & 0.24 & 0.22 \\ 0 & 0.01 & 0.01 & 0.02 & 0.32 & -0.34 & 0.49 & 0.70 & 0.23 & 0.18 \end{pmatrix}$$

The holistic representation for the documents d_1 to d_{10} in a 2-dimensional space is $H := V_2^T$. The columns of both matrices V^T and V_2^T are the representations of the ten documents. In the original space, document d_1 has the holistic representation

$$V^T(d_1) = \left\langle \; 0 \quad 0 \quad 0 \quad 0.01 \quad 0.10 \quad -0.06 \quad -0.18 \quad -0.01 \quad 0.24 \quad 0.22 \; \right\rangle^T ,$$

which is reduced to the lower rank representation

$$V_2^T(d_1) = \left\langle \; 0 \quad 0 \; \right\rangle^T .$$

Consider that the determination of k, the number of singular values to keep, can be regarded as an optimization problem. In the following, we show the results for different k's and discuss the results briefly.

4.1.2 Holistic Information Retrieval

We have seen that a document can be presented as a vector in a k-dimensional space, and accordingly, a k-dimensional query vector can be derived from a string-based query as $\vec{q_k} = \Sigma_k^{-1} U_k^T \vec{q}$ (cf. Equation 3.13 and Equation 3.14). If the query vector $\vec{q_k}$ and the document representation $V^T(d_i)$ have a small distance value $sim(\vec{q}, V^T(d_i)) \leq \theta$, where θ is a threshold, the associated document d_i will be returned to the user.

Sample query for the *Athletics* domain In general, users expect documents about athletics events for the string query "high jump." In the following we compute the latent structure of the repository via LSI and receive the following results for the input query "high jump." The corresponding vector $\vec{q}$ is:

$$\vec{q} = \left\langle \; 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \; \right\rangle^T$$

When the query vector $\vec{q}$ is directly compared against all documents (doc) with the cosine similarity $sim(\vec{q}, V^T(d_i))$ (see Equation 3.15), we receive the result:

doc	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim	0.73	0.50	0.35	0.82	0.84	0.61	0.09	0.31	0.34	0.01

In our methodology, a threshold value θ is an input parameter and it is required in order to determine which documents users will receive.

If $\theta = 0.7$, the document hits will be the documents d_1, d_4, and d_5 which describe athletics events. The documents d_2, d_3, and d_6 also describe athletics events but are not in the result set. If $\theta = 0.35$, the user will receive all athletics documents $(d_1 \ldots d_6)$. However, in general a small threshold value is not a good approach for receiving documents with high precision. Moreover, a user expects especially high jump news (here: document d_1) using the query "high jump" and not all documents about athletics news, for this reason a high threshold value is required. Though, if $\theta = 0.95$, there will be no hits. For the case of an empty result set, the threshold value should be reduced automatically. For the same input query the corresponding query vector in the 2-dimensional space is:

$$\vec{q_2} = \left\langle \begin{array}{cc} -0.02 & 0.23 \end{array} \right\rangle^T$$

If the query vector $\vec{q_2}$ is directly compared against all documents with the cosine similarity $sim(\vec{q_k}, V_k^T(d_i))$, the result will be:

doc	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim	1.00	0.92	1.00	1.00	1.00	1.00	0.94	1.00	1.00	0.01

If $\theta = 0.7$ was the threshold value, the documents hits will be d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, and d_9. In this result set there are all documents about athletics and three fairy tales documents. In order to reduce the false positive rate the threshold value should be increased. If $\theta = 0.95$ was the threshold value, the document hits will be $d_1, d_3, d_4, d_5, d_6, d_8$, and d_9. In this case the documents d_2 and d_7 are no longer in the result set because d_2 is a document about pole vault and d_7 about Prince Charming. We see, that the similarity results of the query vectors ($\vec{q}$ and $\vec{q_2}$) differ for the input query "high jump". The retrieval results using the original space is more precise than using the 2-dimensional space. After presenting a second example for another string query, we will see that for the dimensions $k = 4$ and $k = 5$ the retrieval results are similar to the original space.

Sample query for the $FairyTale$ domain For the query “snow white” the corresponding query vector $\vec{q}$ is:

$$\vec{q} = \left\langle \begin{array}{ccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{array} \right\rangle^T$$

If the query vector $\vec{q}$ is compared against all documents with the cosine similarity approach, the result will be

doc	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim	0.00	0.02	0.00	0.00	0.00	0.00	0.18	0.02	0.01	1.00

If we choose $\theta = 0.95$ as a threshold value, we will retrieve the document d_{10}. The reduced query vector $\vec{q}_2$ is:

$$\vec{q}_2 = \left\langle \begin{array}{cc} 1.40 & 0.20 \end{array} \right\rangle^T$$

If the query vector $\vec{q}_2$ is compared against all documents with the cosine similarity approach, the result will be[1]:

doc	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
sim	-0.04	0.39	-0.03	-0.04	-0.04	-0.03	0.34	0.06	0.0	1.0

If we choose $\theta = 0.95$ as a threshold value, we will retrieve the document d_{10}. This result set is equal to the result set above because the repository only contains one document about Snow White which has very little resemblance to the other documents in the repository. In contrast, the repository contains more than one document about high jump and the similarity of high jump documents to other documents is higher. The next paragraphs show similarity results for the dimension $k = 2$ to $k = 10$ for both input queries: “high jump” and “snow white” with different threshold values.

Query overview for different dimensions We mentioned above that the choice of k is an optimization problem, and we presented results w.r.t. different k’s because for our methodology, k is an input parameter, and we suggest a

[1] The similarity measure “sim” contains negative values because of the cosine function. It is possible to choose a negative threshold value but it is not an intuitive approach. It might indeed be scale the threshold value between 0 and 1.

k based on our experimental results. Table 4.3 shows the document similarity values for the query term “high jump” with the dimensions $k = 2$ to $k = 10$. In the following, we will see that recall and precision results indeed differ and depend on the choice of dimension k and threshold θ.

doc	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
d_1	1	0.97	0.94	0.94	0.92	0.81	0.81	0.73	0.73
d_2	0.92	0.91	0.9	0.93	0.85	0.61	0.51	0.5	0.5
d_3	1	0.96	0.94	0.93	0.43	0.35	0.36	0.35	0.35
d_4	1	0.96	0.93	0.94	0.91	0.84	0.82	0.82	0.82
d_5	1	0.97	0.93	0.93	0.9	0.86	0.84	0.84	0.84
d_6	1	0.95	0.93	0.93	0.86	0.79	0.77	0.61	0.61
d_7	0.94	0.92	0.13	0.09	0.09	0.09	0.09	0.09	0.09
d_8	1	0.71	0.71	0.32	0.31	0.31	0.31	0.31	0.31
d_9	1	0.75	0.76	0.35	0.35	0.35	0.34	0.34	0.34
d_{10}	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.01

Table 4.3: **Document similarity values for the string query “high jump” with the dimensions** $k = 2$ **to** $k = 10$.

In order to find the best choice of k and θ, we have a look to the recall and precision results (see Table 4.4, Table 4.5, and Table 4.6). We would like to mention that in our example the values for recall and precision often is 1 because we have used a small repository for demonstration purposes. For big repositories recall and precision are smaller than 1. In Table 4.4 recall and precision results for the string query “high jump” with the dimensions $k = 2$ to $k = 10$ and the threshold $\theta = 0.95$ are given. Table 4.4 shows that we have high recall and high precision for threshold $\theta = 0.95$ and dimension $k = 2$ or $k = 3$. Otherwise recall and precision is very low.

$\theta = 0.95$	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
recall	0.57	0.8	0	0	0	0	0	0	0
precision	1	1	0	0	0	0	0	0	0

Table 4.4: **Recall and precision results for the string query “high jump” with the dimensions** $k = 2$ **to** $k = 10$**, and the threshold** $\theta = 0.95$.

In Table 4.5 recall and precision results for the string query “high jump” with the dimensions $k = 2$ to $k = 10$ and the threshold $\theta = 0.94$ are given.

$\theta = 0.94$	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
recall	0.5	0.8	0.8	1	0	0	0	0	0
precision	1	1	1	0.5	0	0	0	0	0

Table 4.5: **Recall and precision results for the string query “high jump” with the dimensions** $k = 2$ **to** $k = 10$**, and the threshold** $\theta = 0.94$.

Table 4.5 shows that we have high recall and high precision for threshold $\theta = 0.94$ and the dimensions $k = 2$, $k = 3$, $k = 4$, and $k = 5$. Otherwise recall and precision are very low. More concretely, we have best recall results for $k = 5$, and we have best precision results for $k = 2$, $k = 3$, or $k = 4$. If we have to decide for the best k w.r.t. precision and recall, $k = 3$, $k = 4$, and $k = 5$ will be good candidates.

In Table 4.6 recall and precision results for the string query “high jump” with the dimensions $k = 2$ to $k = 10$, and the threshold $\theta = 0.5$ are given.

$\theta = 0.5$	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
recall	0.44	0.44	0.5	0.67	0.8	0.8	0.8	0.8	0.8
precision	1	1	1	1	1	1	1	1	1

Table 4.6: **Recall and precision results for the string query “high jump” with the dimensions** $k = 2$ **to** $k = 10$**, and the threshold** $\theta = 0.5$.

Table 4.6 shows that we have the best recall and precision results for threshold $\theta = 0.5$ and the dimensions $k = 6$, $k = 7$, $k = 8$, $k = 9$, and $k = 10$.

Figure 4.3 and Figure 4.4 present recall and precision results graphically.

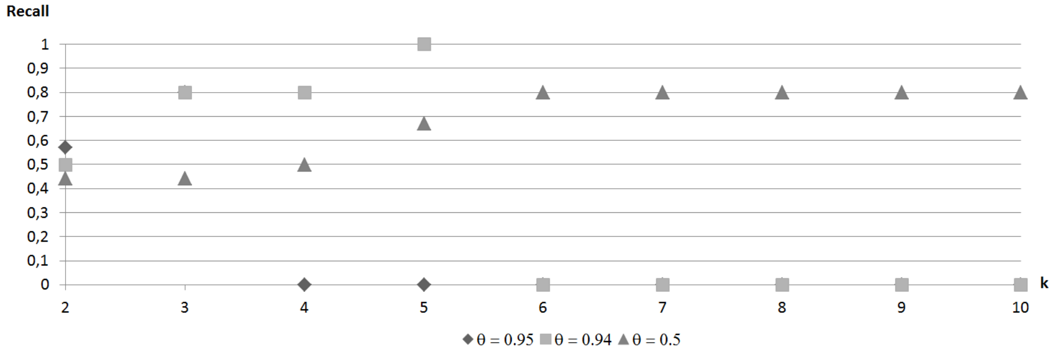


Figure 4.3: Recall results for the string query "high jump" with the dimensions $k = 2$ to $k = 10$ with different thresholds

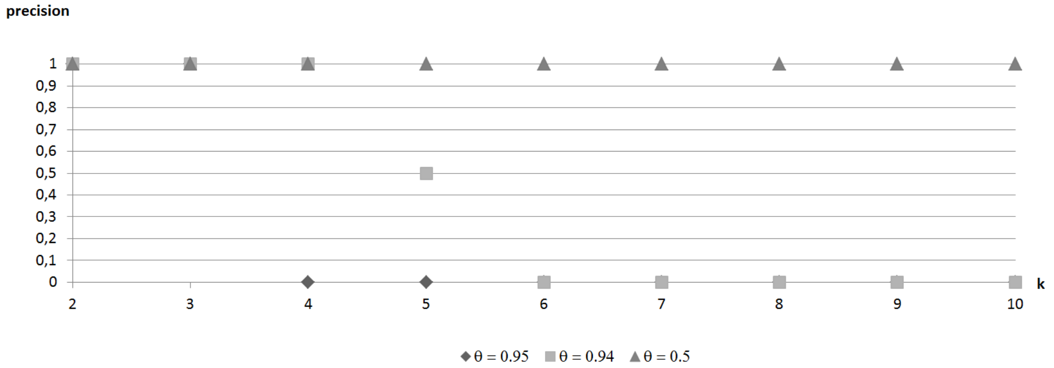


Figure 4.4: Precision results for the string query "high jump" with the dimensions $k = 2$ to $k = 10$ with different thresholds

Taking both diagrams into account, then $k = 4$ and $\theta = 0.94$ are the best parameters for an information retrieval system for our specific example because we have high recall (0.8) and high precision (1.0). For our Snow White example, Table 4.7 shows document similarity values for the string query "snow white" with the dimensions $k = 2$ to $k = 10$.

In order to find the best choice of k and θ for this example, we have a look at the recall and precision results (see Table 4.8). In the following, recall and precision results for the string query "snow white" with the dimensions $k = 2$ to $k = 10$ and the threshold $\theta = 0.95$ are given:

If we choose the threshold values $\theta = 0.94$ or $\theta = 0.5$, we will receive the same recall and precision results shown in Table 4.8. As we have mentioned above, the similarity between the Snow White document d_{10} and the others in

doc	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
d_1	-0.04	-0.04	-0.01	0	0	0	0	0	0
d_2	0.39	0.3	0.23	0.04	0.03	0.03	0.2	0.02	0.02
d_3	-0.03	-0.03	0	0	0	0	0	0	0
d_4	-0.04	-0.03	0	0	0	0	0	0	0
d_5	-0.04	-0.04	0.01	0	0	0	0	0	0
d_6	-0.03	-0.03	0	0	0	0	0	0	0
d_7	0.34	0.33	0.18	0.18	0.18	0.18	0.18	0.18	0.18
d_8	0.06	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02
d_9	0	0	0.01	0.02	0.01	0.01	0.01	0.01	0.01
d_{10}	1	1	1	1	1	1	1	1	1

Table 4.7: **Document similarity for the string query "snow white" with the dimensions $k = 2$ to $k = 10$.**

$\theta = 0.95$	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
recall	1	1	1	1	1	1	1	1	1
precision	1	1	1	1	1	1	1	1	1

Table 4.8: **Recall and precision results for the string query "snow white" with the dimensions $k = 2$ to $k = 10$, and the threshold $\theta = 0.95$.**

the repository is very small. Consequently the result set is very precise. But usually the challenge is that a document cannot be easily distinguished from other documents while only considering used terms in the documents. For our knowledge management scenario we have IR results with high recall and high precision if we set the parameters $k = 4$ and $\theta = 0.94$, so that the best holistic representation is $H = V_4^T$.

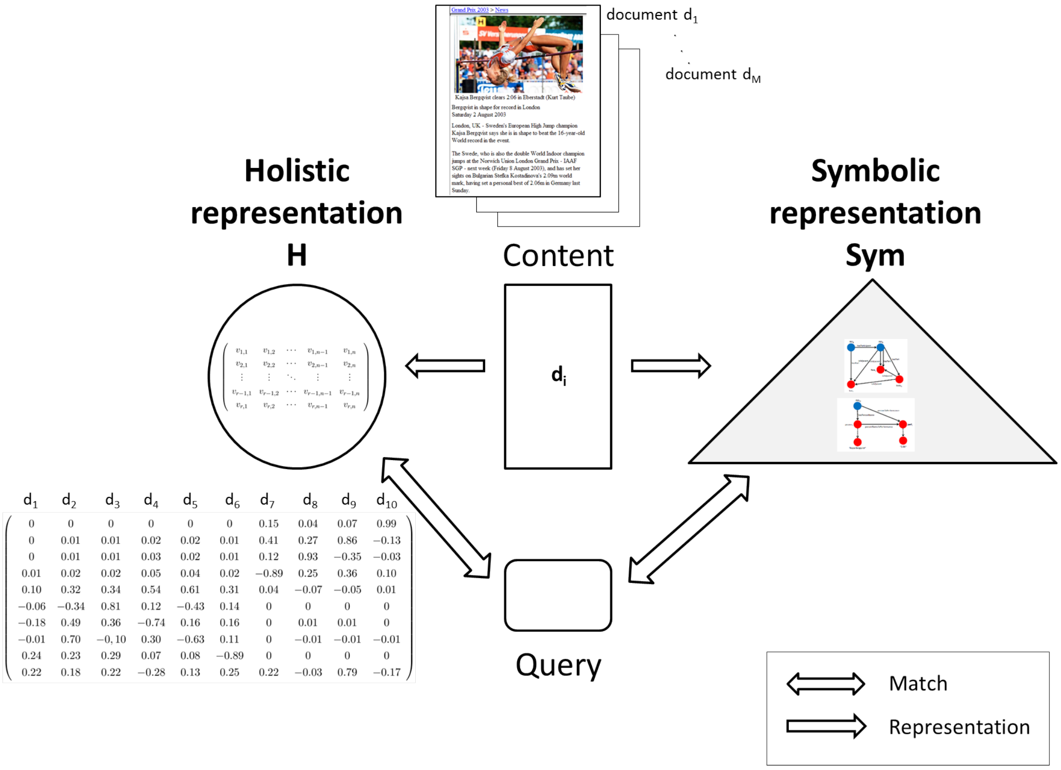
In order to find the best dimension for a holistic representation in general, can be solved by using an IR system with a feedback process. A feedback process is characterized by a learning process in which the retrieved result set is evaluated by user's surf behavior. If a user visits a web site which is in the result set, this web site is classified as a hit. On the basis of this information, the best k can be found. But the best choice of k and threshold θ do not imply to have the best recall and precision results (cf. Table 4.4). The result set using the parameters $k = 4$ and $\theta = 0.94$ is a good compromise w.r.t recall and precision but this is just a compromise. In this thesis, we will present how to increase recall and at least maintaining precision while systematically combining holistic and symbolic approaches. Another approach for evaluating test result is to use plot a *receiver operating characteristic* (ROC) curve[2] with the true positive against the false positive rate. The true positive rate is also known as recall. False positive rate is also known as fall-out. The accurancy is measured under the ROC curve. An area of 1 represents a perfect test and an area of 0.5 represents a worthless test.[3] Plots of such curves are useful for interpreting medical test results as well as IR result. But in the context of information retrieval precision and recall values are used as an alternative to ROC.

[2] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[3] http://gim.unmc.edu/dxtests/roc3.htm

4.2 Symbolic Content Description Approach

In this section we present a symbolic content description approach for our knowledge management scenario presented in Section 2.2.2. For the purpose of increasing precision we use description logics as a language for presenting symbolic representations *Sym* for the content of each document d_i (see Figure 4.5 [right]).



$$\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n-1} & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{r-1,1} & v_{r-1,2} & \cdots & v_{r-1,n-1} & v_{r-1,n} \\ v_{r,1} & v_{r,2} & \cdots & v_{r,n-1} & v_{r,n} \end{pmatrix}$$

$$\begin{array}{cccccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 & d_{10} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.04 & 0.07 & 0.99 \\ 0 & 0.01 & 0.01 & 0.02 & 0.02 & 0.01 & 0.41 & 0.27 & 0.86 & -0.13 \\ 0 & 0.01 & 0.01 & 0.03 & 0.02 & 0.01 & 0.12 & 0.93 & -0.35 & -0.03 \\ 0.01 & 0.02 & 0.02 & 0.05 & 0.04 & 0.02 & -0.89 & 0.25 & 0.36 & 0.10 \\ 0.10 & 0.32 & 0.34 & 0.54 & 0.61 & 0.31 & 0.04 & -0.07 & -0.05 & 0.01 \\ -0.06 & -0.34 & 0.81 & 0.12 & -0.43 & 0.14 & 0 & 0 & 0 & 0 \\ -0.18 & 0.49 & 0.36 & -0.74 & 0.16 & 0.16 & 0 & 0.01 & 0.01 & 0 \\ -0.01 & 0.70 & -0,10 & 0.30 & -0.63 & 0.11 & 0 & -0.01 & -0.01 & -0.01 \\ 0.24 & 0.23 & 0.29 & 0.07 & 0.08 & -0.89 & 0 & 0 & 0 & 0 \\ 0.22 & 0.18 & 0.22 & -0.28 & 0.13 & 0.25 & 0.22 & -0.03 & 0.79 & -0.17 \end{array}$$

Figure 4.5: Representation types: holistic and symbolic content descriptions with the focus on "symbolic representation".

In this thesis the process for creating symbolic representations *Sym* for each d_i is called symbolic knowledge creation process (SKCP) and is illustrated in Figure 4.6. The three main processes are *analysis process*, *interpretation process*, and *fusion process*.

Analysis process The SKCP diagram shows that the analysis process enriches objects from the pre-iconographical level by giving a meaning to such

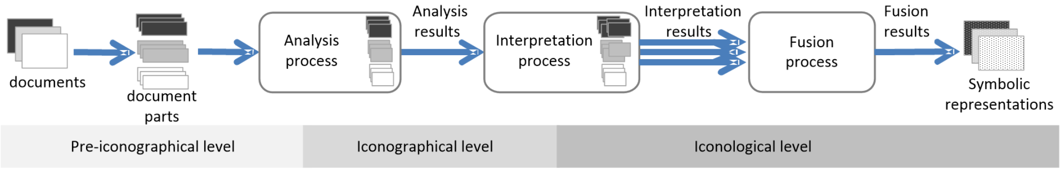


Figure 4.6: Symbolic Knowledge Creation Process (SKCP) with the main processes analysis process, interpretation process, and fusion process for creating symbolic representations for each document.

objects (iconographical level). More technically, in the analysis process documents will be annotated by so-called *low-level* extraction tools (IE) with the aim to find annotations which represent content of a document manually or automatically. Low-level IE tools such as DeepDive, NELL (Never-Ending Language Learning), or M-OntoMat Annotizer which identify objects and relations among the objects within documents, videos, and HTML pages are presented in [MCH+15, CBK+10, Pal09].

Interpretation process The interpretation process computes interpretations for the analysis result set with the aim to enrich symbolic content descriptions delivered by the analysis process that means there is a switch from the iconographical to the iconological level. In contrast to Panofsky's definition, here, the interpretation process can also deliver many interpretations for one document part.

Fusion process The fusion process merges interpretation results given from different media parts such as image, caption, and text from the same document. The union of symbols for image, caption, and text aims at a symbolic description for a document. The level of description does not change after the execution of fusion.

4.2.1 Symbolic Knowledge Creation Process

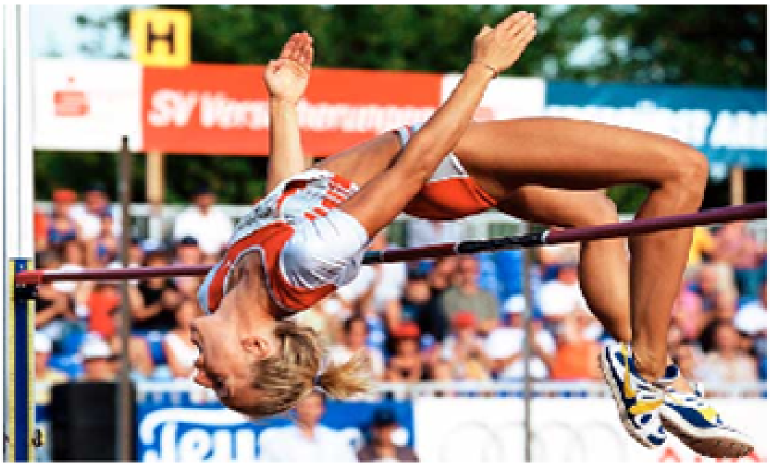
The symbolic knowledge creation process needs documents as input (see Figure 4.6). In our example we use documents from different web sites such as International Association of Athletics Federations (IAAF) and Wikipedia as representative repositories in order to demonstrate the creation process for symbolic representations and symbolic information retrieval.

In the project BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction), which was a research project funded by the European Union under the Information Society Technologies program (IST-FP6-027538), a symbolic content description creation process was developed for creating symbolic representations for the athletics domain. The document content descriptions are stored in a so-called *BOEMIE repository.* In this repository representations are directly linked to the corresponding document via an URL. We use this BOEMIE repository and Wikipedia pages in order to demonstrate our methodology suggested in this work (see Chapter 5). The symbolic representations for the Wikipedia pages had to be created for our symbolic IR purposes. Therefore the service Calais [CFO10] was used. Calais automatically extracts semantic information from web pages, in other words, it reads unstructured text and classified entities, facts, and events within a text.

In the following we present the three processes analysis, interpretation and fusion process of the symbolic knowledge creation process for a better understanding how symbolic representations are created, which are used for our knowledge management scenario. As a representative example we present how symbolic representations are generated for the document parts (image, caption, and text) from document d_1. The other symbolic representations for documents *Docs* from the BOEMIE repository ($\{d_2 \ldots d_6\} \subseteq Docs$) were created in the same way. The documents from the $FairyTale$ domain ($\{d_7 \ldots d_{10}\} \subseteq Docs$) were annotated by the service Calais. The specific A-box for each document $d_7 \ldots d_{10} \subseteq Docs$ is presented in Appendix A. Consider that the quality of annotations differs by using different annotation services.

Analysis process

The document d_1 from the BOEMIE repository presented in Figure 4.2 has the document parts: image and caption (see Figure 4.7). Textual information in the caption supplements visual information in the image by providing additional information such as the athlete's name, performance, and city. It is assumed that the multimedia document in Figure 4.7 has successfully been partitioned into image, caption, and text parts before the analysis process continues.



Kajsa Bergqvist clears 2:06 in Eberstadt (Kurt Taube)

Figure 4.7: **Image and caption from Figure 4.2.** Source: IAAF [IAA09].

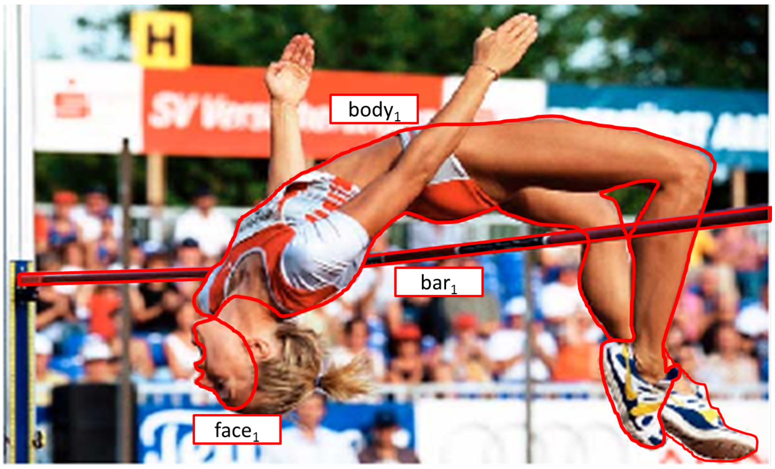


Figure 4.8: **Annotated image.** Low-level objects in the image: bar_1, $body_1$, $face_1$, and $image_1$. Image source: IAAF [IAA09].

In the analysis process the following objects are extracted for the image in Figure 4.8: bar_1, $body_1$, $face_1$, and $image_1$; and the relations $isAdjacent(body_1, face_1)$ and $isAdjacent(body_1, bar_1)$. In addition, the four objects are linked with concepts, and stored in modality-specific A-boxes (image: see Figure 4.9 and caption: see Figure 4.10). The analysis A-box for the image in Figure 4.7 is presented in Figure 4.9. The A-box contains the assertions:

$domain_1$	:	*Athletics*
$image_1$	:	*Image*
bar_1	:	*HorizontalBar*
$body_1$	:	*PersonBody*
$face_1$	:	*PersonFace*
$(body_1, face_1)$	:	*isAdjacent*
$(body_1, bar_1)$	:	*isAdjacent*

Figure 4.9: **The analysis A-box AnalysisAboxImage.** This A-box represents the results of image analysis for the image in Figure 4.7.

- $domain_1$: *Athletics*, which represents the domain of this document part,

- $image_1$: *Image*, which represents the modality of this document part,

- $face_1$: *PersonFace*, bar_1 : *HorizontalBar*, and $body_1$: *PersonBody*, which represent objects in the image at the iconographical level, and

- $(body_1, face_1)$: *isAdjacent*, $(body_1, bar_1)$: *isAdjacent*, which represent relations of objects in the image.

$domain_1$	:	*Athletics*
$caption_1$	:	*Caption*
$pname_1$	:	*PersonName*
$perf_1$	:	*Performance*
$(pname_1, perf_1$	:	*personNameToPerformance*
($pname_1$, “Kajsa Bergqvist”)	:	*hasValue*
($perf_1$, “2.06”)	:	*hasValue*

Figure 4.10: **The analysis A-box AnalysisAboxCaption.** This A-box represents the results of caption analysis for the caption in Figure 4.7.

The A-box in Figure 4.10 contains the assertions:

- $domain_1$: *Athletics*, which represents the domain of this document part,

- $caption_1$: *Caption*, which represents the modality of this document part,

- $pname_1$: *PersonName* and $perf_1$: *Performance*, which represent objects in the caption at the iconographical level,

- $(pname_1, perf_1)$: *personNameToPerformance*, which represents a relation in the caption, and

- ($pname_1$, “Kajsa Bergqvist”) :*hasValue*, ($perf_1$, “2.06”): *hasValue*, which represent attributes in the caption.

For the text part in document d_1 (Figure 4.2) the associated A-box is presented in Figure 4.11.

$domainsn_1$	:	*Athletics*
$text_1$	:	*Text*
sn_1	:	$SportsName$
$(pn_1, perf_1)$	:	$personNameToPerformance$
$(pn_2, perf_2)$	:	$personNameToPerformance$
$city_1$	:	*City*
($city_1$, “London”)	:	$hasCityNameValue$
$event_1$	:	$SportsEventName$
($event_1$, “Norwich Union London Grand Prix”)	:	$hasSportsEventNameValue$
hjn_1	:	$HighJumpName$
(hjn_1, “High Jump”)	:	$hasSportsNameValue$
pn_1	:	$PersonName$
(pn_1, “Kajsa Bergqvist”)	:	$hasPersonNameValue$
pn_2	:	$PersonName$
$perf_1$	:	$Performance$
($perf_1$, “2.06”)	:	$hasPerformanceValue$
$perf_2$	:	$Performance$
($perf_2$, “2.09”)	:	$hasPerformanceValue$
d_1	:	*Date*
(d_1, “Friday 8 August 2003”)	:	$hasStartDateValue$

Figure 4.11: **Analysis A-box** ***Text*****.** This A-box represents the results of text analysis for the text in Figure 4.2.

Analogously to both A-boxes described above, this A-box contains the assertions:

- $domainsn_1$: *Athletics*, which represent the domain of this document part,

- *text*$_1$: *Text*, which represents the modality of this document part,

- *sn*$_1$: $SportsName$, *city*$_1$: *City*, $event_1$: $SportsEventName$, hjn_1 : $HighJumpName$, *pn*$_1$: $PersonName$, *pn*$_2$: $PersonName$, $perf_1$: $Performance$, $perf_2$: $Performance$, d_1 : *Date*, which represent objects in the caption at the iconographical level,

- (*pn*$_1$,$perf_1$) : $personNameToPerformance$ and (*pn*$_2$,$perf_2$) :*person-*$NameToPerformance$ represent the relations between the assertions *pn*$_1$: *PersonName* and $perf_1$: $Performance$, *pn*$_2$: *PersonName* and $perf_2$: $Performance$ within the text, and

- (*city*$_1$, “London”) : *hasCity*$NameValue$, ($event_1$, “Norwich Union London Grand Prix”): *hasSports*$EventNameValue$, (hjn_1, “High Jump”) : *hasSports*$NameValue$, (*pn*$_1$, “Kajsa Bergqvist”) : *hasPerson*$Name$-$Value$, ($perf_1$, “2.06”) : $hasPerformanceValue$, ($perf_2$, “2.09”): *has-*$PerformanceValue$, (d_1, “Friday 8 August 2003”):*hasStartDate*$Value$, which represent attributes given by the text.

The three A-boxes are symbolic representations of document d_1 at the preiconographical level. The interpretation process delivers representations at the iconological level. This process is described as follows.

Interpretation process

The interpretation process computes interpretations for the analysis results in this work by using the *A-box abduction* algorithm in order to enrich symbolic content descriptions delivered by the analysis process. We call this kind of interpretation process high-level interpretation process. As described in Section 3.3 the *abduction* process is used to find explanations (causes) for observations (effects). In our case observations are for example symbolic content descriptions for images, texts, and videos. As described in Section 3.3.2, abduction is formalized as

$$\mathcal{KB} \cup \Delta \models \Gamma,$$

where the background knowledge ($\mathcal{KB}$), and observations (Γ) are given and explanations (Δ) are to be computed. For example, the A-boxes presented in Figure 4.9, Figure 4.10, and Figure 4.11 are observations. To compute the explanation Δ in our context we modify this equation into

$$\mathcal{KB} \cup \Gamma_1 \cup \Delta \models \Gamma_2, \tag{4.1}$$

where the assertions in Γ will be split into bona fide assertions (Γ_1) and assertions requiring fiats (Γ_2). Bona fide assertions are assumed to be believed to be true by default, whereas fiat assertions are aimed to be explained. The abduction process tries to find explanations (Δ) such that Γ_2 is entailed. This entailment decision on the abduction process is implemented as (boolean) query answering. The output Δ of the abduction process represents the enhanced symbolic content descriptions. Multiple solutions are possible. Consequently, a ranking of explanation is needed. We rank explanations via Equation 3.31 in order to receive a preferred Δ.

In this work the interpretation process is done by a semantic interpretation service called BIWS (BOEMIE Interpretation Web Services) [Kay11, Chapter 4.2]. BIWS is part of the BOEMIE system and is a high-level extraction tool which is offered by a semantic interpretation engine. BIWS supports the interpretation by implementing the A-box abduction operation (more details of the abduction process are described formally in Section 3.3.1). To this end it uses the inference services provided by RacerPro[4]. As an output, BIWS generates interpretation data based on the following input: symbolic annotations, T-box and rules (see Ontologien.zip[5]).

For instance, A-box abduction is assumed to be used on the observations image and caption analysis Aboxes (Figure 4.9, and Figure 4.10), the T-box in Figure 4.12, and rules in Figure 4.13 as input. The T-box contains intentional knowledge in the form of a terminology (see Figure 4.12). Sports specific rules and T-boxes are defined during the BOEMIE project which are called Athletics Event Ontology (AEO), where all concepts and relations regarding the athletics domain are modeled, the Multimedia Content Ontology (MCO), which has been defined to address structural aspects of multimedia content, and the Geographic Information Ontology (GIO), where notions for representing geographic information are modeled.

[4] www.racer-systems.com

[5] https://www.ifis.uni-luebeck.de/fileadmin/user_files/ifis/files/melzer/Ontologien.zip

$$\begin{array}{rcl}
\mathit{Athletics} \sqcap \mathit{FairyTale} & \sqsubseteq & \bot \\
\mathit{Person} & \sqsubseteq & \mathit{Human} \\
\mathit{Pole} & \sqsubseteq & \mathit{SportsEquipment} \\
\mathit{Bar} & \sqsubseteq & \mathit{SportsEquipment} \\
\mathit{Pole} \sqcap \mathit{Bar} & \sqsubseteq & \bot \\
\mathit{JumpingEvent} & \sqsubseteq & \exists_{\geq 1} hasParticipant.Person \\
\mathit{PoleVault} & \sqsubseteq & \mathit{JumpingEvent} \sqcap \exists hasPart.HorizontalBar \sqcap \\
 & & \exists hasPart.Pole \\
\mathit{HighJump} & \sqsubseteq & \mathit{JumpingEvent} \sqcap \exists hasPart.HorizontalBar \\
\ldots & &
\end{array}$$

Figure 4.12: **An excerpt of $\mathcal{KB}$ consisting of a T-box $\mathcal{T}$.**

$$\begin{array}{lcl}
\mathit{isAdjacent}(Y,Z) & \leftarrow & \mathit{Person}(X), \mathit{hasPart}(X,Y), \mathit{PersonFace}(Y), \\
 & & \mathit{hasPart}(X,Z), \mathit{PersonBody}(Z) \\
\mathit{isAdjacent}(Y,Z) & \leftarrow & \mathit{PoleVault}(X), \mathit{hasPart}(X,Y), \mathit{HorizontalBar}(Y), \\
 & & \mathit{hasPart}(X,W), \mathit{Pole}(W), \mathit{hasParticipant}(X,Z), \\
 & & \mathit{PoleVaulter}(Z) \\
\mathit{isAdjacent}(Y,Z) & \leftarrow & \mathit{HighJump}(X), \mathit{hasPart}(X,Y), \mathit{HorizontalBar}(Y), \\
 & & \mathit{hasParticipant}(X,Z), \mathit{HighJumper}(Z) \\
\ldots & &
\end{array}$$

Figure 4.13: **An excerpt of rules.**

In order to find explanations for image and caption, the A-box Γ is divided into Γ_1 (bona fide assertions) and Γ_2 (fiat assertions) following Equation 4.1. In this example, the bona fide assertions of analysis A-box *AnalysisAboxImage* presented in Figure 4.9 are:

$$\Gamma_{1(image)} = \{\mathit{Athletics}(\mathit{domain}_1), \mathit{Image}(\mathit{image}_1), \mathit{HorizontalBar}(\mathit{bar}_1), \\ \mathit{PersonBody}(\mathit{body}_1), \mathit{PersonFace}(\mathit{face}_1)\}.$$

Γ_1 has all assertions from the analysis A-box *AnalysisAboxImage* without the role assertions. And Γ_2 contains the fiats:

$$\Gamma_{2(image)} = \{\mathit{isAdjacent}(\mathit{body}_1, \mathit{face}_1), \mathit{isAdjacent}(\mathit{body}_1, \mathit{bar}_1)\}.$$

Γ_2 has all role assertion from the analysis A-box *AnalysisAboxImage.* The bona fide assertions for the analysis A-box *AnalysisAboxCaption* presented in Figure 4.10 are:

$$\begin{aligned}\Gamma_{1(caption)} = \{&Athletics(domain_1), Caption(caption_1),\\ &PersonName(pname_1), Performance(perf_1),\\ &hasValue(pname_1, \text{Kajsa Bergqvist}),\\ &hasValue(perf_1, 2.06)\}.\end{aligned}$$

Here Γ_1 has the same assertions from the A-box *AnalysisAboxCaption* presented in Figure 4.10. Γ_2 is:

$$\Gamma_{2(caption)} = \{personNameToPerformance(pname_1, perf_1)\}.$$

Γ_2 has the one role assertion from the analysis A-box *AnalysisAboxCaption.*

The DL reasoner RacerPro provides the function *retrieve-with-explanation*, which is an implementation of the A-box abduction algorithm. The function *retrieve-with-explanation* accepts a strategy parameter that defines the strategy for instantiating variables. There are two possible values: "use new individuals" and "reuse existing individuals" (:reuse-old). If the *retrieve-with-explanation* function is called without the optional strategy parameter, the value is "use new individuals", and thus the function prefers to hypothesize new individual names instead of reusing existing individual names while generating explanations. The *retrieve-with-explanation* function can also be instructed to additionally generate explanations where existing individual names are reused. If the function *retrieve-with-explanation* is called with the optional parameter value reuse-old, which corresponds to the value "reuse existing individuals", it tries to reuse existing individual names as part of an explanation, if such individual names exist in the A-box.

For the queries

$$Q_{(image)} := \{()|isAdjacent(body_1, face_1)\}$$

$$Q_{(image_2)} := \{()|isAdjacent(body_1, bar_1)\}$$

which contain the role assertions of $\Gamma_{2(image)}$, the exact syntax of the RacerPro function calls are:

```
((retrieve-with-explanation ()
 (body1 face1 isadjacent) (:reuse-old))
```

```
((retrieve-with-explanation ()
 (body1 bar1 isadjacent) (:reuse-old))
```

Assume that the three rules for query expansion presented in Figure 4.13 are given. All rules have the atom *isAdjacent* in the head, and thus can be exploited to generate explanations for the boolean query $Q_{(image)}$:

$$\begin{aligned}
\Gamma_{2.1(image)} = \{&Person(IND_1), hasPart(IND_1, face_1), PersonFace(face_1),\\
&hasPart(IND_1, body_1), PersonBody(body_1)\}\\
\Gamma_{2.2(image)} = \{&PoleVault(IND_1), hasPart(IND_1, face_1), HorizontalBar(face_1),\\
&hasPart(IND_1, IND_2), Pole(IND_2), PoleVaulter(body_1),\\
&hasParticipant(IND_1, body_1)\}\\
\Gamma_{2.3(image)} = \{&HighJump(IND_1), hasPart(IND_1, face_1), HorizontalBar(face_1),\\
&hasParticipant(IND_1, body_1), HighJumper(body_1)\}
\end{aligned}$$

The *retrieve-with-explanation* function returns explanations, which contains the set of non-entailed assertions from $\Gamma_{2.1}$, $\Gamma_{2.2}$, and $\Gamma_{2.3}$:

$$\Delta_{1(image)} = \{Person(IND_1), hasPart(IND_1, face_1), hasPart(IND_1, body_1)\}.$$

The result set represents that *PersonBody*(*body₁*) and *PersonFace*(*face₁*) are entailed from Γ_1.

$$\begin{aligned}
\Delta_{2(image)} = \{&HighJump(IND_6), HighJumper(IND_1), isAdjacent(IND_1, bar_1),\\
&hasPart(IND_6, bar_1), hasParticipant(IND_6, IND_1)\}.
\end{aligned}$$

This result set represents that $Person(IND_1)$, $HorizontalBar(bar_1)$, $Person-Body(body_1)$, $PersonFace(face_1)$, $isAdjacent(IND_1, face_1)$, $isAdjacent(body_1, face_1)$, $hasPart(IND_1, face_1)$, $hasPart(IND_1, body_1)$, and $isAdjacent(body_1, bar_1)$ are entailed from Γ_1.

$$\begin{aligned}
\Delta_{3(image)} = \{&PoleVault(IND_6), IND_1 : PoleVaulter, isAdjacent(IND_1, bar_1),\\
&hasPart(IND_6, bar_1), hasParticipant(IND_6, IND_1)\}.
\end{aligned}$$

This result set represents that $Person(IND_1)$, $PersonBody(body_1)$, $Person-Face(face_1)$, $HorizontalBar(bar_1)$, $isAdjacent(IND_1, face_1)$, $isAdjacent(body_1, face_1)$, $hasPart(IND_1, face_1)$, $hasPart(IND_1, body_1)$, and $isAdjacent(body_1, bar_1)$ are entailed from Γ_1. However, only $\Gamma_{2.1}$ is consistent w.r.t. $\mathcal{T}$ and $\mathcal{A}$ because of some disjointness axioms, i.e. $PoleVault$ and $HighJump$ are disjoint.

In order to retrieve an interpretation A-box, during the interpretation process consistent assertions are added to the analysis A-box. Therefore, the rules are applied in a forward-chaining way by using the so-called *execute-or-reexecute-all-rules* function. After executing the rules, in this example the consistent assertions of $\Delta_{1(image)}$ are added to the A-box *AnalysisAboxImage* from Figure 4.9. A new interpretation A-box for the analysis A-box *AnalysisAboxImage* is presented in Figure 4.14.

$domain_1$	:	$Athletics$
$image_1$	:	$Image$
bar_1	:	$HorizontalBar$
$body_1$	:	$PersonBody$
$face_1$	:	$PersonFace$
$(body_1, face_1)$	:	$isAdjacent$
$(body_1, bar_1)$	:	$isAdjacent$
$\mathbf{IND}_1$	:	$\mathbf{Person}$
$(\mathbf{IND}_1, \mathbf{face}_1)$	:	$\mathbf{hasPart}$
$(\mathbf{IND}_1, \mathbf{body}_1)$	:	$\mathbf{hasPart}$

Figure 4.14: **A new interpretation A-box InterpretationAboxImage$_1$.** The analysis A-box presented in Figure 4.9 is added with $\Delta_{1(image)}$.

The *retrieve-with-explanation* function delivers for next query $Q_{(image_2)}$ no answers, because all explanations that can be generated are inconsistent.

Kaya argues in [Kay11] that there might be different levels of interpretations. A level i is the number of recursive calls of the interpretation process. In the beginning of an interpretation process the fiats have level 0. In our example the A-box presented in Figure 4.14 has interpretation results at level 0. The choice of an appropriate level is discussed in Kaya's thesis. For demonstrating the interpretation process by example, we choose level 1 because the A-box in Figure 4.14 represents objects at the pre-iconographical level and symbolic interpretations at level 1 represents objects at the iconographical level as we will see in the following.

At level 1, $\Gamma^1_{2(image)}$ contains the fiat:

$$\Gamma^1_{2(image)} = \{isAdjacent(IND_1, bar_1)\},$$

where the power of Γ denotes the interpretation level.

Analogously, the *retrieve-with-explanation* function delivers two further (preferred) explanations $\Delta_{2(image)}$ and $\Delta_{3(image)}$ which are presented in Figure 4.15 and Figure 4.16.

$domain_1$	:	*Athletics*
$image_1$	:	*Image*
IND$_6$	:	**HighJump**
IND$_1$	:	**HighJumper**
IND_1	:	*Person*
bar_1	:	*HorizontalBar*
$body_1$	:	*PersonBody*
$face_1$	:	*PersonFace*
(IND_1, bar_1)	:	*isAdjacent*
(IND_6, bar_1)	:	*hasPart*
(IND_6, IND_1)	:	*hasParticipant*
$(IND_1, face_1)$	:	*isAdjacent*
$(body_1, face_1)$	:	*isAdjacent*
$(IND_1, face_1)$	:	*hasPart*
$(IND_1, body_1)$	:	*hasPart*
$(body_1, bar_1)$	:	*isAdjacent*

Figure 4.15: **The interpretation A-box InterpretationAboxImage$_2$**. The analysis A-box in Figure 4.9 is added with $\Delta_{2(image)}$.

$domain_1$	:	*Athletics*
$image_1$	:	*Image*
IND$_6$	:	**PoleVault**
IND$_1$	:	**PoleVaulter**
IND_1	:	*Person*
bar_1	:	*HorizontalBar*
$body_1$	:	*PersonBody*
$face_1$	:	*PersonFace*
(IND_1, bar_1)	:	*isAdjacent*
(IND_6, bar_1)	:	*hasPart*
(IND_6, IND_1)	:	*hasParticipant*
$(IND_1, face_1)$	:	*isAdjacent*
$(body_1, face_1)$	:	*isAdjacent*
$(IND_1, face_1)$	:	*hasPart*
$(IND_1, body_1)$	:	*hasPart*
$(body_1, bar_1)$	:	*isAdjacent*

Figure 4.16: **The interpretation A-box InterpretationAboxImage$_3$.** The analysis A-box in Figure 4.9 is added with $\Delta_{3(image)}$.

In order to reduce the number of explanations, a scoring function (see Equation 3.31) as described in Section 3.3.2 is used. Explanation with the highest scores are in the result set. In our example the scores are:

$$S(\Delta_{2(image)}) = S_i(\Delta_{2(image)}) - S_h(\Delta_{2(image)}) = |6+1| - |6+2| = 7 - 8 = -1$$
$$S(\Delta_{3(image)}) = S_i(\Delta_{3(image)}) - S_h(\Delta_{3(image)}) = |6+1| - |6+2| = 7 - 8 = -1$$

Both explanations $\Delta_{2(image)}$ and $\Delta_{3(image)}$ have the same score in this example. It is a plausible result because the image in document d_1 could not exactly interpreted as a high jump image because detected objects in the image are not enough for finding an interpretation which represents high jump objects for an image which represents a high jumper. In RacerPro, the preference score is implemented accordingly and returns, in our example, $\Delta_{2(image)}$ and $\Delta_{3(image)}$.

We know from our knowledge that document d_1 describes high jump events and the high jumper "Kajsa Bergqvist" and that therefore $\Delta_{2(image)}$ is the "right" explanation, or, to put in other words $\Delta_{3(image)}$ reduces precision. We now present an approach how to find explanation results with high precision.

In [Kay11] is shown that fusion of modalities within a document can increase precision. In the following we demonstrate the fusion process after creating an interpretation representation for the caption part of document d_1.

For a better understanding of the fusion process, the Figure 4.17 illustrates the A-box *InterpretationAboxImage*$_2$ (see Figure 4.15) (without the individual names *domain*$_1$ and *image*$_1$). The red nodes demonstrate the analysis instance assertions and the blue nodes represent the new instance assertions after the interpretation process.

The particular A-box for the caption part of document d_1 is the A-box *AnalysisAboxCaption* which is presented in Figure 4.10. During the interpretation process the following query is used with the input of $\Gamma_{2(caption)}$:

$$Q_{(caption)} := \{() \mid personNameToPerformance(pname_1, perf_1)\}.$$

The exact syntax of the RacerPro function call is for reusing existing individuals (:reuse-old):

```
((retrieve-with-explanation ()
 (pname1 perf1 personnametoperformance) (:reuse-old))
```

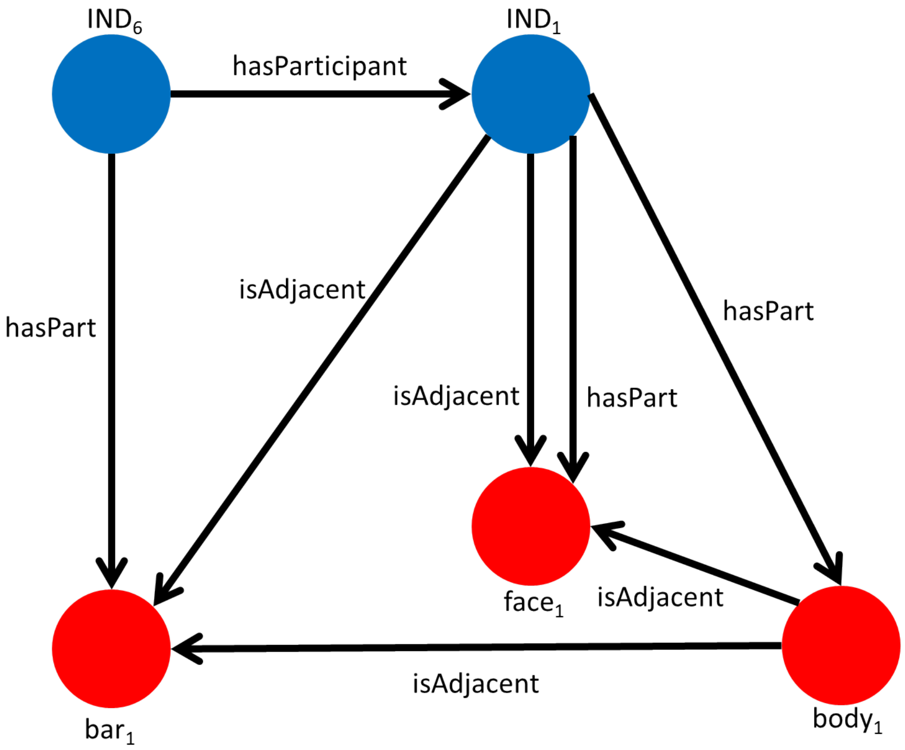


Figure 4.17: **Illustration of the interpretation A-box InterpretationAboxImage$_2$:** Red nodes demonstrate the analysis assertions (bottom), blue nodes demonstrate the new assertions after the interpretation process (top).

The *retrieve-with-explanation* function delivers one explanation:

$$\Delta_{1(caption)} = \{Person(IND_{11}), hasPersonName(IND_{11}, pname_1), \\ personToPerformance(IND_{11}, perf_1)\}.$$

The specific interpretation A-box *InterpretationAboxCaption* with $\Delta_{1(caption)}$ is presented in Figure 4.18. In addition Figure 4.19 illustrates the interpretation process (without the nodes $domain_1$ and $caption_1$). The red nodes demonstrate the analysis assertions and the blue nodes represent the new created assertions. In contrast to the image interpretation A-box, the caption interpretation A-box contains attribute assertions, namely $hasValue(pname_1$, “Kajsa Bergqvist”) and $hasValue(perf_1$, “2.06”). The interpretation process is also done for the textual part of document d_1. One interpretation A-box for the textual part is presented in Figure 4.20. For the fusion process, the symbolic representations of the three modalities image, caption, and text for d_1 is presented next.

$$\begin{array}{rcl}
\mathit{domain}_1 & : & \mathit{Athletics} \\
\mathit{caption}_1 & : & \mathit{Caption} \\
\mathit{pname}_1 & : & \mathit{PersonName} \\
\mathit{perf}_1 & : & \mathit{Performance} \\
\mathbf{IND}_{11} & : & \mathbf{Person} \\
(\mathbf{IND}_{11}, \mathbf{pname}_1) & : & \mathbf{hasPersonName} \\
(\mathbf{IND}_{11}, \mathbf{perf}_1) & : & \mathbf{personToPerformance} \\
(\mathit{pname}_1 \mathit{perf}_1) & : & \mathit{personNameToPerformance} \\
(\mathit{pname}_1, \text{“Kajsa Bergqvist”}) & : & \mathit{hasValue} \\
(\mathit{perf}_1, \text{“2.06”}) & : & \mathit{hasValue}
\end{array}$$

Figure 4.18: **InterpretationAboxCaption**. The analysis A-box *AnalysisAboxCaption* presented in Figure 4.10 extended with $\Delta_{1(caption)}$.

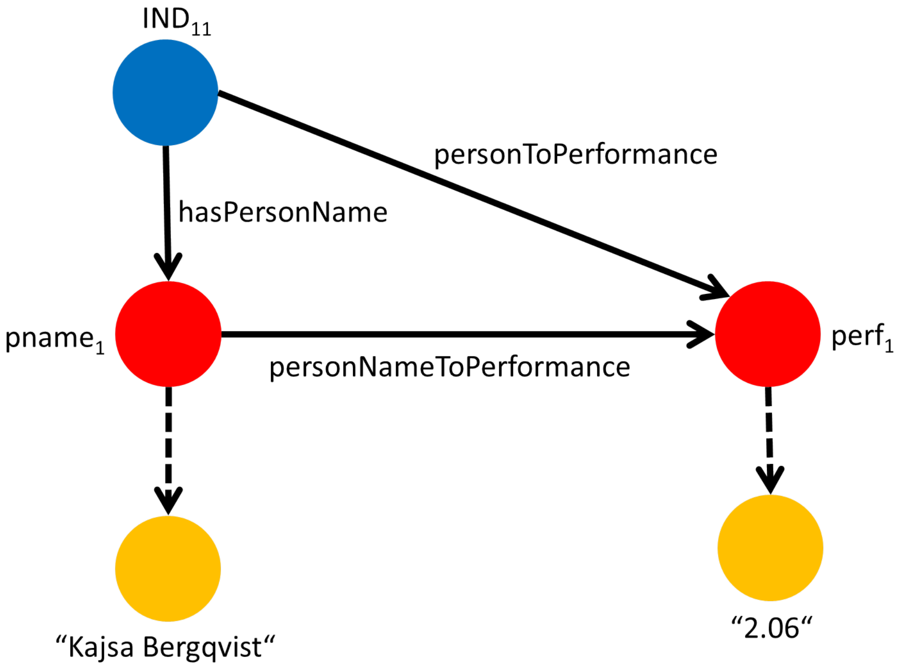


Figure 4.19: **Illustration of the interpretation A-box InterpretationAboxCaption:** Red nodes demonstrate the analysis assertions, orange nodes are attribute assertions (bottom), and blue nodes demonstrate the new assertions after the interpretation process (top).

Fusion process

Multimedia documents such as web pages are partitioned into segments possibly w.r.t. different modalities (image, caption, and text) as we have seen above.

domain_1	:	*Athletics*
text_1	:	*Text*
IND_8	:	*HighJump*
IND_{51}	:	*Person*
IND_{50}	:	*Person*
sn_1	:	*SportsName*
$(\mathit{IND}_{50}, \mathrm{perf}_1)$	:	*personToPerformance*
$(\mathit{IND}_{50}, \mathrm{pn}_1)$	:	*hasPersonName*
$(\mathit{IND}_{51}, \mathrm{pn}_2)$	:	*hasPersonName*
$(\mathit{IND}_{51}, \mathrm{perf}_2)$	:	*personToPerformance*
$(\mathit{pn}_1, \mathrm{perf}_1)$	:	*personNameToPerformance*
$(\mathit{pn}_2, \mathrm{perf}_2)$	:	*personNameToPerformance*
city_1	:	*City*
(city_1,“London”)	:	*hasCityNameValue*
event_1	:	*SportsEventName*
(event_1, “Norwich Union London Grand Prix”)	:	*hasSportsEventNameValue*
hjn_1	:	*HighJumpName*
(hjn_1, “High Jump”)	:	*hasSportsNameValue*
pn_1	:	*PersonName*
(pn_1, “Kajsa Bergqvist”)	:	*hasPersonNameValue*
pn_2	:	*PersonName*
perf_1	:	*Performance*
(perf_1, “2.06”)	:	*hasPerformanceValue*
perf_2	:	*Performance*
(perf_2, “2.09”)	:	*hasPerformanceValue*
d_1	:	*Date*
(d_1, “Friday 8 August 2003”)	:	*hasStartDateValue*

Figure 4.20: **Interpretation A-box *Text*.** This A-box represents the interpretation results for the text in Figure 4.2.

Modality-specific analysis and interpretation processes are then applied to each segment to obtain modality-specific interpretation A-boxes. We have seen that a document may even have multiple interpretations: image, caption, and text interpretations. The challenge is to collate the three modality-specific interpretations to a single coherent representation for a document, in other words, the idea is to fuse all interpretations with the aim to increase the precision of IR results. Fusion means that two individual names i, j which refer to the same object are identified with *same-as*(i, j). In [Kay11, EKM09a, EKM09b] the authors present how to fuse interpretation results for a multimedia document with structure-oriented rules created by experts. In this work we present two fusion approaches: one fusion approach with using structure-oriented rules and one without these rules. Fusion is a core element for our new methodology, which is presented in Chapter 5.

Fusion process with structure-oriented rules In [Kay11] it is described that each A-box has a modality-specific instance assertion of the particular media type. This means that all analysis and interpretation A-boxes have the assertion $image_1$: *Image*, $caption_1$: *Caption*, or $text_1$: *Text*. In addition, all modality-specific individual names ($image_1$, $caption_1$, and $text_1$) are set to a new artificial *depicts* relation (e.g., $depicts(image_1, IND_6)$). The new relations are essential for the structure-oriented rule approach, so that the following rules can be used for fusing image, caption, and text A-boxes:

$$\begin{aligned} hasCaption(X, A) \leftarrow & Image(X), depicts(X, Y), Caption(A), \\ & depicts(A, B), same\text{-}as(Y, B) \end{aligned} \tag{4.2}$$

$$\begin{aligned} hasImage(X, A) \leftarrow & Text(X), depicts(X, Y), Image(A), \\ & depicts(A, B), same\text{-}as(Y, B) \end{aligned} \tag{4.3}$$

$$\begin{aligned} hasText(X, A) \leftarrow & Caption(X), depicts(X, Y), Text(A), \\ & depicts(A, B), same\text{-}as(Y, B) \end{aligned} \tag{4.4}$$

The idea of *same-as* assertions is that individuals names will be fused when rules are used abductively (with backward-chaining as explained above). For fusing image, caption, and text interpretation A-boxes presented in our example, the interpretation A-boxes have new *depicts* assertions. We present in the following how to fuse image and caption interpretation A-boxes. Both A-boxes with artificial *depicts* relations (see Figure 4.21 and Figure 4.18).

For creating same-as assertion for the purpose of fusing image and caption interpretation A-boxes, the fusion rule *hasCaption* (see Equation 4.2) has to be executed, then Γ_2 contains the fiat assertion:

$$\Gamma_2 = \{hasCaption(image_1, caption_1)\},$$

$$\begin{array}{|rcl|}
\hline
\mathit{domain}_1 & : & \mathit{Athletics} \\
\mathit{image}_1 & : & \mathit{Image} \\
\mathit{IND}_6 & : & \mathit{HighJump} \\
\mathit{IND}_1 & : & \mathit{HighJumper} \\
\mathit{IND}_1 & : & \mathit{Person} \\
\mathit{bar}_1 & : & \mathit{HorizontalBar} \\
\mathit{body}_1 & : & \mathit{PersonBody} \\
\mathit{face}_1 & : & \mathit{PersonFace} \\
(\mathit{IND}_1, \mathit{bar}_1) & : & \mathit{isAdjacent} \\
(\mathit{IND}_6, \mathit{bar}_1) & : & \mathit{hasPart} \\
(\mathit{IND}_6, \mathit{IND}_1) & : & \mathit{hasParticipant} \\
(\mathit{IND}_1, \mathit{face}_1) & : & \mathit{isAdjacent} \\
(\mathit{body}_1, \mathit{face}_1) & : & \mathit{isAdjacent} \\
(\mathit{IND}_1, \mathit{face}_1) & : & \mathit{hasPart} \\
(\mathit{IND}_1, \mathit{body}_1) & : & \mathit{hasPart} \\
(\mathit{body}_1, \mathit{bar}_1) & : & \mathit{isAdjacent} \\
(\mathbf{image}_1, \mathbf{IND}_6) & : & \mathbf{depicts} \\
(\mathbf{image}_1, \mathbf{IND}_1) & : & \mathbf{depicts} \\
(\mathbf{image}_1, \mathbf{bar}_1) & : & \mathbf{depicts} \\
(\mathbf{image}_1, \mathbf{body}_1) & : & \mathbf{depicts} \\
(\mathbf{image}_1, \mathbf{face}_1) & : & \mathbf{depicts} \\
\hline
\end{array}$$

Figure 4.21: **The interpretation A-box InterpretationAboxImage from Figure 4.15 extended with *depicts* assertions.**

which is translated to the query:

$$Q := \{()|\mathit{hasCaption}(\mathit{image}_1, \mathit{caption}_1)\}.$$

The explanation is:

$$\Delta_{1(\text{image, caption})} = \{\mathit{same\text{-}as}(\mathit{IND}_1, \mathit{IND}_{11})\}.$$

The result shows that the individual name IND_1 is equal to the individual name IND_{11}.

$domain_1$	:	$Athletics$
$caption_1$	:	$Caption$
$pname_1$	:	$PersonName$
$perf_1$	:	$Performance$
IND_{11}	:	$Person$
$(IND_{11}, pname_1)$	:	$hasPersonName$
$(IND_{11}, perf_1)$	:	$personToPerformance$
$(pname_1 perf_1)$	:	$personNameToPerformance$
$(pname_1$, "Kajsa Bergqvist")	:	$hasValue$
$(perf_1$, "2.06")	:	$hasValue$
$(\mathbf{caption}_1, \mathbf{pname}_1)$	:	**depicts**
$(\mathbf{caption}_1, \mathbf{perf}_1)$	:	**depicts**
$(\mathbf{caption}_1, \mathbf{IND}_{11})$	:	**depicts**

Figure 4.22: **The A-box InterpretationAboxCaption extended with *depicts* assertions.**

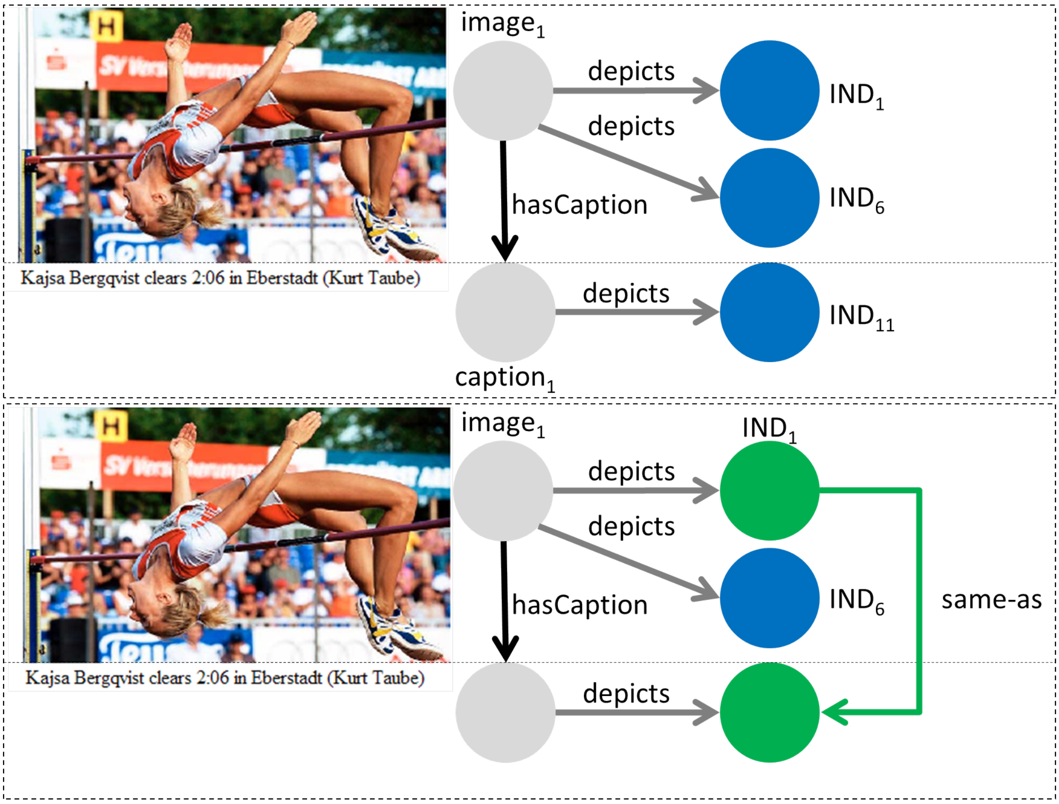


Figure 4.23: **Illustration of the result of a fusion process using rules.** In the beginning (top): IND_1 from image interpretation A-box and IND_{11} of caption interpretation A-box are not fused. After the fusion process (bottom): IND_1 and IND_{11} are fused via *same-as* assertion so that image and caption assertions are associated.

Figure 4.23 illustrates the result of a fusion process: In the beginning, objects from different modalities are not fused, e.g., such as the individual names IND_1 from the image interpretation A-box and IND_{11} from the caption interpretation A-box, where the modalities are from the same document d_1. After the fusion process, the individuals IND_1 and IND_{11} (green nodes) are fused. In Section 4.2.2 it is described that the fusion of image and caption interpretation A-boxes has advantages w.r.t. precision. However, fusion using structure-oriented rules has the disadvantage that the number of individuals increases drastically, and therefore the performance for symbolic IR purposes is quite low. The reason is that every instance assertion in an A-box has n new created depicts role assertions and, in addition, the corresponding number of queries is also n, and n queries have to be answered by symbolic IR systems. An evaluation of the fusion process with structure-oriented rules versus the new fusion process using the A-box difference operator (see below) is given in Section 5.2.

Fusion process using the A-box difference operator A fusion algorithm working without using structure-oriented rules is called *A-box fusion* and is presented in Algorithm 1. The algorithm needs a T-box $\mathcal{T}$, and two A-boxes $\mathcal{A}$ and $\mathcal{B}$ as input. In our example, the A-box $\mathcal{A}$ is the interpretation A-box for the image in document d_1 and A-box $\mathcal{B}$ is the interpretation A-box for the caption part in document d_1. The T-box $\mathcal{T}$ is presented in Figure 4.12 and is used for demonstrating how the A-box difference operator works. The A-box difference operator *abox_diff* is used for computing the mappings $\phi(\mathcal{A})$ and $\phi(\mathcal{B})$ as well as differences $\Delta_{\mathcal{A},\mathcal{B}}$ and $\Delta_{\mathcal{B},\mathcal{A}}$ w.r.t. the T-box $\mathcal{T}$. The differences help to identify the commonalities of $\mathcal{A}$ and $\mathcal{B}$. If there are two individual names $i : \mathcal{A}, j : \mathcal{B}$ which refer to the same object, they will be identified with $\textit{same-as}(i, j)$. The *same-as* assertions are added to an A-box $\mathcal{C}$. A fusion A-box $\mathcal{F}$ is the symbolic representation for a document and is computed via $\mathcal{F} := A \cup B \cup \mathcal{C}$. The *A-box fusion* returns a fused A-box $\mathcal{F}$.

We give an example: the A-boxes $\textit{InterpretationAboxImage}_2$ (see Figure 4.15) and *InterpretationAboxCaption* (see Figure 4.18) are fused as follows: RacerPro offers the function *compute-abox-difference* which is used for computing the $(\phi(\textit{Caption}), \Delta_{\textit{Image},\textit{Caption}})$ and $(\phi(\textit{Image}), \Delta_{\textit{Caption},\textit{Image}})$.

Algorithm 1 The *A-box fusion* algorithm.

```
function $fusion(\mathcal{T}, \mathcal{A}, \mathcal{B})$ :
  $\mathcal{F} := \emptyset$ //Fusion A-box
  $\mathcal{C} := \emptyset$ //A-box with equal assertions
  $(\Delta_{\mathcal{A},\mathcal{B}}, \phi(\mathcal{B})) = abox\_diff(\mathcal{T}, \mathcal{A}, \mathcal{B})$
  $(\Delta_{\mathcal{B},\mathcal{A}}, \phi(\mathcal{A})) = abox\_diff(\mathcal{T}, \mathcal{B}, \mathcal{A})$
  if $(i \in \phi(\mathcal{A}) \cup\ j \in \phi(\mathcal{B})$, where $i, j$ refer to the same object) then
    $\mathcal{C} := \mathcal{C} \cup \{same\text{-}as(i, j)\}$
  end if
  $\mathcal{F} := A \cup B \cup \mathcal{C}$
  return $\mathcal{F}$
```

The returned results are:

$$\begin{aligned}\Delta_{Image,Caption} = \{ & HorizontalBar(bar_1), isAdjacent(IND_1, bar_1), \\ & HighJumper(IND_1), hasParticipant(IND_6, IND_1), \\ & HighJump(IND_6), hasPart(IND_6, bar_1), \\ & PersonFace(face_1), isAdjacent(IND_1, face_1), \\ & PersonBody(body_1), hasPart(IND_1, face_1), \\ & Image(image_1), isAdjacent(body_1, face_1), \\ & hasPart(IND_1, body_1)\}\end{aligned}$$

because $\phi(Caption)$ has a mapping $\{IND_{11} \mapsto IND_1\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{IND_1 : Person\})) \models \{IND_{11} : Person\}$.

$$\begin{aligned}\Delta_{Caption,Image} = \{ & personToPerformance(IND_{11}, perf_1), \\ & PersonName(pname_1), Performance(perf_1), \\ & hasPersonName(IND_{11}, pname_1), \\ & personNameToPerformance(pname_1, perf_1), \\ & Caption(caption_1)\}\end{aligned}$$

because $\phi(Image)$ has a mapping $\{IND_1 \mapsto IND_{11}\}$, since there is the entailed assertion $(\mathcal{T}, \phi(\{IND_{11} : Person\})) \models \{IND_1 : Person\}$.

IND_1 and IND_{11} refer to the same object and therefore set to the *same-as* assertion *same-as*(IND_1, IND_{11}). The specific fusion A-box $\mathcal{F}$ is presented in Figure 4.24.

$domain_1$	:	$Athletics$
$image_1$	:	$Image$
IND_6	:	$HighJump$
IND_1	:	$HighJumper$
IND_1	:	$Person$
bar_1	:	$HorizontalBar$
$body_1$	:	$PersonBody$
$face_1$	:	$PersonFace$
(IND_1, bar_1)	:	$isAdjacent$
(IND_6, bar_1)	:	$hasPart$
(IND_6, IND_1)	:	$hasParticipant$
$(IND_1, face_1)$	:	$isAdjacent$
$(body_1, face_1)$	:	$isAdjacent$
$(IND_1, face_1)$	:	$hasPart$
$(IND_1, body_1)$	:	$hasPart$
$(body_1, bar_1)$	:	$isAdjacent$
$domain_1$	:	$Athletics$
$caption_1$	:	$Caption$
$pname_1$	:	$PersonName$
$perf_1$	:	$Performance$
IND_{11}	:	$Person$
$(IND_{11}, pname_1)$	:	$hasPersonName$
$(IND_{11}, perf_1)$	:	$personToPerformance$
$(pname_1 perf_1)$	:	$personNameToPerformance$
$(pname_1$, “Kajsa Bergqvist”$)$	:	$hasValue$
$(perf_1$, “2.06”$)$	:	$hasValue$
$(\mathbf{IND}_1, \mathbf{IND}_{11})$	:	**same-as**

Figure 4.24: **Symbolic representation** Sym_{d_1} **for document** d_1**.**

This example shows how to fuse A-boxes without structure-oriented rules. If in the A-box $InterpretationAboxImage_2$ we had two instance assertions which refer to the same object ($Person$), assume $Person(IND_1)$ and $Person(IND_4)$, and in the A-box there only one instance assertion which refers to the object Person $Person(IND_{11})$, it could happen that the A-box $\mathcal{C}$ contained the two assertions: $same\text{-}as(IND_1, IND_{11})$ and $same\text{-}as(IND_4,\ IND_{11})$. In this case, the second assertion is not desired. To this end a-box fusion approach decreases false positives but increases the false negatives. Fusing of same modalities via a-box fusion, this approach is only applicable if the T-box has been adequately specified or if image and caption interpretation A-boxes only contain one concept name at a time. In Section 5.2 an evaluation of the a-box difference operator is given.

4.2.2 Symbolic Information Retrieval

In the following, we present the advantages and disadvantages of both fusion algorithms for symbolic information retrieval tasks.

With respect to a T-box $\mathcal{T}$ and a symbolic content description Sym_i of document d_i, an online symbolic query answering problem *answers* for retrieving relevant documents is defined as (cf. Section 3.3.1):

$$answers((\mathcal{T}, Sym_i), cq) \tag{4.5}$$

where cq is an A-box query.

Example 1 Assume an image of multimedia document $d_1 \in Docs$ (see Figure 4.2) with the associated symbolic representations presented in Figure 4.15 and Figure 4.16 enriched with *depicts* role assertions. Further assume an engineer who is interested in documents with images representing **high jump** events and uses the symbolic query:

$$cq_{hj} := \{(x, y) | Image(x), HighJump(y), depicts(x, y)\}.$$

The information retrieval service RacerPro delivers the answer

$$\{(image_1, IND_1)\}$$

because $image_1$ and IND_1 are parts of the A-boxes (symbolic content descriptions of d_1). The A-boxes are linked to the documents via URIs, and accordingly the engineer retrieves document d_1 as a result.

Example 2 Assume an image of multimedia document $d_1 \in \mathit{Docs}$ (see Figure 4.2) with the associated symbolic representations presented in Figure 4.15 and Figure 4.16 enriched with *depicts* role assertions. And assume another engineer is interested in images with **pole vault** events and uses the symbolic query:

$$cq_{pv} := \{(x, y) | \mathit{Image}(x), \mathit{PoleVault}(y), \mathit{depicts}(x, y)\}.$$

The answer of RacerPro is:

$$\{(\mathit{image}_1, \mathit{IND}_1)\}.$$

Document d_1 will be returned as a result for cq_{hj} and cq_{pv} because the symbolic interpretations for the image in document d_1 are also interpreted as a pole vault event. The reason is that the image in document d_1 could not exactly be interpreted as a pole vault or a high jump image by the interpretation process, because for the image in document d_1 there exist two symbolic representations presented in Figure 4.15 and Figure 4.16. To this end, there are two symbolic content descriptions for d_1, and the second engineer retrieves document d_1 as a false positive document.

If we have a fused symbolic representation, presented in Figure 4.24, for document d_1, RacerPro will deliver an empty set for cq_{pv} because the symbolic representation A-box of d_1 represents a high jump event.

Precision and recall Suppose we have a document repository represented by a set of documents $\mathit{Docs} = \langle d_1, \ldots, d_{10} \rangle$ and by a symbolic representation Sym_{doc} associated with each document *doc*. For the purpose of comparability with the holistic approach we define a set of queries in which people could be interested in choosing documents from the *Athletics* or *FairyTale* domain. The queries are

$$\begin{aligned}
cq_1 :=& \{x | \mathit{HighJump}(x)\}\\
cq_2 :=& \{x | \mathit{PoleVault}(x)\}\\
cq_3 :=& \{x | \mathit{FairyTale}(x)\}\\
cq_4 :=& \{x | \mathit{Person}(x), \mathit{hasPersonName}(\text{“Snow White”}, x)\}\\
cq_5 :=& \{x | \mathit{Person}(x), \mathit{hasPersonName}(\text{“Kajsa Bergqvist”}, x)\}
\end{aligned}$$

Tables 4.9 and 4.10 represent the retrieval results for queries cq_1 to cq_5 versus documents d_1 to d_{10} with symbolic content descriptions generated by analysis or interpretation processes. Hits are marked with ✓and no hits with ✗.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
cq_1	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
cq_2	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗
cq_3	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
cq_4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
cq_5	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗

Table 4.9: **Information retrieval results using symbolic content descriptions for multimedia content representations (analysis).**

For the queries cq_1, cq_2, and cq_5, the symbolic IR system delivers the high jump document d_1. It would be desirable not to obtain d_1 if we want to receive pole vault documents (cq_2). This example shows that symbolic IR systems deliver documents with low precision.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
cq_1	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
cq_2	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗
cq_3	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
cq_4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
cq_5	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗

Table 4.10: **Information retrieval results using symbolic content descriptions for multimedia content representations (interpretation).**

In Table 4.10 we see the information retrieval results for all symbolic content descriptions (*interpretation*). There is one difference (highlighted in green). The symbolic IR system delivers two hits for the queries cq_1, cq_2, and cq_5. For queries cq_1 and cq_5 the retrieval systems delivers the high jump document d_1 as a result. We see that the interpretation process leads to an increase of precision. Recall and precision results are given in Tables 4.11 and 4.12.

	cq_1	cq_2	cq_3	cq_4	cq_5
recall	1	1	1	1	1
precision	1	0.6	1	1	1

Table 4.11: Recall and precision results for queries cq_1 to cq_5 (analysis).

	cq_1	cq_2	cq_3	cq_4	cq_5
recall	1	1	1	1	1
precision	1	0.75	1	1	1

Table 4.12: Recall and precision results for queries cq_1 to cq_5 (interpretation).

The recall and precision results in Table 4.11 and Table 4.12 show for a small document set that we have high recall and also high precision for our knowledge management scenario. As a result we confirm with our computations the results presented in Kaya [Kay11] that symbolic representations at the iconological level lead to high precision results (see Figure 4.25).

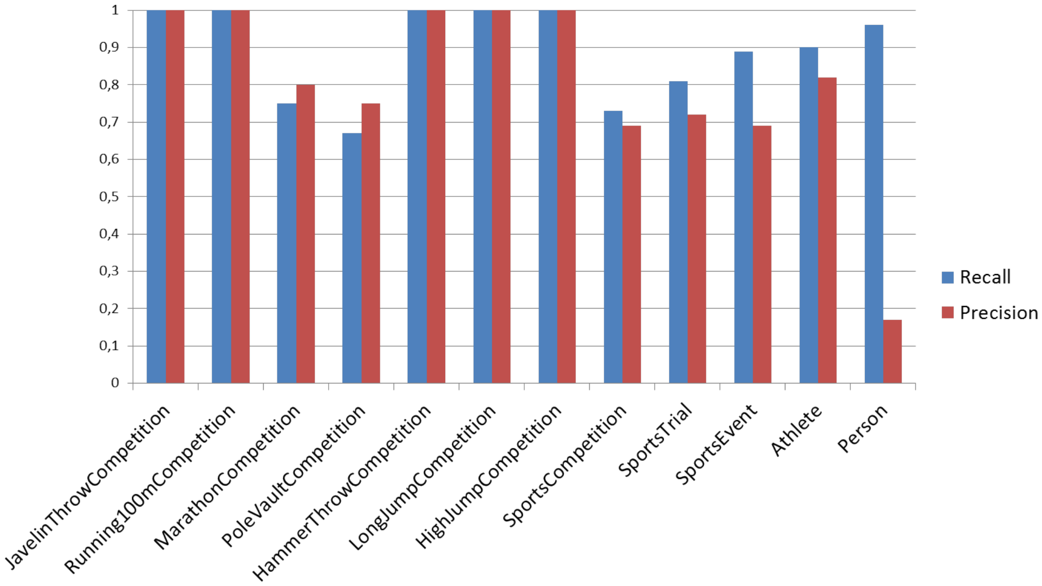


Figure 4.25: **Recall and precision results for using interpretation A-boxes as symbolic representations for documents.** Source for the values: Kaya [Kay11]

Figure 4.25 shows the average recall and precision values for an experiment with 100 documents from the BOEMIE repository. The experimental study delivers good results w.r.t. recall and precision. The concept name *Person* is an outlier and represents that *Person* is denoted in various ways during the analysis process.

In contrast to Kaya's experiments in which only athletics documents are used, our repository contains athletics as well as fairy tale documents for demonstrating the advantages of symbolic IR. We have shown that precision results (here for cq_2) are even better with symbolic content descriptions by using more than one repository. In addition, we have shown that in contrast to the holistic information retrieval results, we do not need any threshold values for receiving precise results. Nevertheless the queries have to be written by experts and symbolic IR tools have to be used.

Chapter 5

Systematic Combination of Holistic and Symbolic Content Descriptions

In the following we present a methodology for the systematic combination of symbolic and holistic content descriptions for information retrieval which we call *HolSym Methodology*. As presented in Chapter 2, fundamental aspects of knowledge management have to be formalized in order to realize Nonaka's knowledge creation process and to provide an automated knowledge management process. In addition, we address non-functional issues such as the performance of the HolSym Methodology and the quality of the results obtained. As an important contribution of this thesis, we would like to explore holistic and symbolic information retrieval in combination, and evaluate the results in a representative knowledge management scenario. We show the evaluation results after the presentation of the proposed methodology. How this methodology can usefully contribute to an automated knowledge management process using so-called semantic assets as *reified knowledge units*, which represent context-specific knowledge, is described at the end of this chapter.

5.1 HolSym Methodology

We have discussed that, on the one hand, holistic methods tend to provide high recall and low precision, and on the other hand, symbolic methods can be characterized by low recall and high precision. The systematic combination of different problem solving methods could have the advantage that in a combined formalism they benefit from each other. In this thesis we suggest a so-called HolSym Methodology for presenting how to surpass or alleviate the limitation of holistic methods by systematically combining holistic and symbolic methods.

5.1.1 Symbolic versus Holistic Information Retrieval

In this section, we give an overview of symbolic and holistic information retrieval. In general the objective of retrieval systems is that users receive high-quality documents. Therefore a query answering problem has to be solved in a way that a user does not receive false positive or false negative documents. For this reason, in the research area of distributional semantics, algorithms were developed, e.g. LSI [DDF$^+$90a], word2vec [MCCD13], Knowledge Vault [DMG$^+$14], DeepDive [Zha15], and NELL [MCH$^+$15] which built latent structures of knowledge in order to provide the delivering of high-quality documents.

LSI, for instance, is an approach to automatic indexing and retrieval. It solves the problem to match the strings in a user's query with those of relevant documents. If words have multiple meanings, or documents are written with words which equal to an implicit knowledge representation, an estimation of a so-called semantic space is required for solving query answering problems with high recall and precision. LSI has a statistical technique to compute the latent structure in this semantic space. This latent structure represents a holistic representation of documents in a repository. To this end we use LSI for computing a holistic representation and use it as a basis for systematically combining holistic and symbolic approaches. Holistic and symbolic query answering algorithms are described in the following before we present the combined IR approach.

Suppose a repository $\mathcal{R}$ is computed offline. $\mathcal{R}$ is represented by a set of documents $Docs = \langle d_1, \ldots, d_n \rangle$, by a holistic representation H, by a symbolic representation (an A-box Sym_{doc}) associated with each document doc, and also by a feature-based representation (FB_{doc}) associated with each document (cf. Figure 1.1), formally:

$$\mathcal{R} := (Docs, H, \mathcal{T}, Sym, FB). \tag{5.1}$$

With respect to a repository of this kind, an online query answering problem QA for retrieving relevant documents is defined as:

$$QA(Q, \mathcal{R}, \theta), \tag{5.2}$$

where Q is a query vector $\vec{q}$ (see Equation 3.2.1) or an A-box query cq (as defined in Section 3.3.1), and θ is a threshold value. Above, we presented two approaches for solving query answering problems, namely holistic and symbolic approaches. Queries referring to FB are well understood and will not be considered here. Nevertheless, FB representation and query language such as Metalog[1] could easily be integrated in our systematic combination of holistic and symbolic representations.

The query answering algorithm for *holistic* IR is defined in Algorithm 2. As discussed in Section 4.1, a holistic document representation V^T is computed as an approximation of a term-document matrix C by one of lower rank using the singular value decomposition (SVD). The document representation $H := V_k^T$ with $V_k^T = \langle doc_1, \ldots, doc_N \rangle$ is a new representation for each document in the collection. Queries will also be cast into the same low-rank representation which are represented by $\vec{q_k}$. The documents $Docs$, their holistic representations presented with H, and the query vector $\vec{q_k}$ are input parameters of Algorithm 2. For solving the query answering problem QA_{hol}, query-document similarity scores via cosine similarity are computed. It means that all documents will be compared with the query vector which is very time-consuming. LSH solves this performance problem. In an implementation of the HolQuery algorithm we suggest to use LSH. For a simple demonstration of our methodology we use the cosine similarity measure. If the query vector $\vec{q_k}$ and the document representation doc_i have a small distance, i.e. the predicate $sim(\vec{q_k}^T, doc_i) \geq \theta$, where θ is a threshold, the associated document *Docs(i)* of doc_i will be in the result set *docs*.

[1] `http://www.w3.org/TandS/QL/QL98/pp/metalog.html`

The query answering algorithm for *symbolic* IR is defined in Algorithm 3. A conjunctive query cq, documents $Docs$, a T-box $\mathcal{T}$, and symbolic representations of the documents Sym_{doc} are input parameters of the *SymQuery* algorithm. The algorithm solves the query answering problem $answers((\mathcal{T}, Sym_{doc}), cq)$. If the query is answered with true, the document $doc \in Docs$ will be in the result set $docs$. Then the result set will be returned to the user.

Algorithm 2 The *HolQuery* algorithm.

```
$QA_{\mathrm{hol}}(\vec{q_k}, (Docs, H, \_, \_, \_), \theta)$:
  $docs := \emptyset$
  for $i = 1$ to $N$ do
    if $sim(\vec{q_k}^T, doc_i \in H_i) \geq \theta$ then
      $docs := docs \cup \{(doc_i, Docs(i))\}$
    end if
  end for
  return $docs$
```

Algorithm 3 The *SymQuery* algorithm.

```
$QA_{\mathrm{sym}}(cq, (Docs, \_, \mathcal{T}, Sym, \_), \_)$:
  $docs := \emptyset$
  for $doc \in Docs$ do
    $ans := answers((\mathcal{T}, Sym_{doc}), cq)$
    if $ans \neq \emptyset$ then
      $docs := docs \cup \{(doc, ans)\}$
    end if
  end for
  return $docs$
```

In Chapter 4, we presented the limitation of holistic and symbolic approaches and showed that the symbolic approach leads to higher precision by combining Information Retrieval approaches.

5.1.2 Combined Information Retrieval

In general it is challenging to combine holistic and symbolic approaches with the aim to increase recall and precision simultaneously. In the context of statistical relational learning many researcher contributions present an highly specialized scientific background for combined IR, e.g., for learning from low-dimensional embeddings [MCCD13], identify a set of plausible formulas from knowledge bases [WMC14], as well as learning latent and distributional representations of Horn clauses to enhance logic-based completion for large datasets [WC16] by using a scalable probabilistic logic called ProPPR [WMC13] in order to build intelligent IR systems to deal with the uncertainty of data representations. Towards a combined IR a standard boolean model was developed [Sal83, SM86]. Nevertheless the potential of systematically combining holistic and symbolic for receiving high-quality documents is far from exhausted.

Clustering approaches such as, e.g., k-means and GVM[2] (Greedy Variance Minimization) clustering can be formulated as an optimization problem for identifying similar groups (clusters) of documents automatically so that the retrieval is effective and efficient (the clustering approach used in this thesis is described below). However for classical clustering algorithms it is difficult to determine clusters in high-dimensional data. In [Rac08], Race proposes a clustering approach via dimension reduction. Dimension reduction is done with LSI in order to reduce the dimension of data so that the inherent clusters become clearer. Race also presents that the accuracy is increased by using LSI in combination with clustering approaches.

In this thesis we pick up the idea to combine LSI with a clustering approach in order to benefit from its improvements. GVM is one clustering algorithm which has been popularly used in the area of IR. If GVM is used via LSI, the two problems have to be solved:

1. The document dimensionality has to be reduce in a way that more semantic of documents is captured.

2. Discover the initial central location of the cluster that can represent most semantic information.

We have presented that symbolic representations capture more precise semantics of documents than holistic ones. In order to solve the first problem, we define a new semantic representation H_{Sem} based on symbolic data which is described below. Then LSI is used to reduce the document dimensionality. For

[2] `http://www.tomgibara.com/clustering/fast-spatial/`

solving the second problem we use the score values delivered by LSI as input parameters for discovering the initial central location of the cluster. These solutions are parts of our new suggested methodology. Therefore a repository $\mathcal{R}'$ is required:

$$\mathcal{R}' := (\mathcal{R}, H_{Sem})$$

such that H_{Sem} is a semantic representation. As we will see in the following, in addition $Docs'$ as a partitioning of $Docs$ of the form $\langle\{d_i \ldots\} \ldots\rangle$ and the associated symbolic representation Sym' for a set of documents in the same cluster are required. The symbolic representation Sym' is a union A-box and represents a *synopsis* of all document representations which are in the same cluster. Symbolic query answering is then accomplished with $QA'_{\text{sym}}(cq, \mathcal{R}, \theta)$. $QA'_{\text{sym}}(.)$ takes care of using sets of documents rather than single documents. How to compute QA'_{sym} is described in the following. In addition, the systematic combination of holistic and symbolic approaches and the required computation of $Docs'$ and Sym' in an online process is presented with a simplified example.

Document Clustering

In Chapter 5.1 the recall and precision results have shown that document representations are not easy to separate, and patterns of document structures are not obvious to find, and therefore sometimes a non-negligible set of outliers are in the result set. Finding outliers can lead to increase precision of IR result sets. Clustering is a way to identify outliers. Clustering is a way to identify outliers by grouping documents (or their representations) such that documents in the same document group g_{rel} are more similar to each other than to those in other groups. In addition, by applying clustering techniques outliers can be identified, which leads to decrease of false positive or true negative rates.

Algorithm 4 The *clusterDocuments* algorithm.

```
$clusterDocuments(\vec{q_k}, H, g)$:
  $S := scores(\vec{q_k}, H)$
  $g_{rel} := GVM(g, S)$
  **return** $(g_{rel}, S)$
```

Algorithm 4 presents the *clusterDocuments* algorithm. A query vector $\vec{q_k}$, a holistic representation H of documents, and the number of clusters (groups) g are input parameters. For an input query vector $\vec{q}$ the similarity values (scores) between query vector $\vec{q_k}$ and document vectors H are computed using the operation $S = scores(\vec{q_k}, H)$ (see Algorithm 5). In a next step the scores are used to find the same groups of documents using the GVM algorithm $g_{rel} = GVM(g, S)$. As a result, each document is associated to a group number.

Algorithm 5 *scores*

```
$scores(\vec{q_k}, H)$:
  $n := size(H)$
  $res :=$ ARRAY$[1 \dots n]$ OF $\mathbb{R}$
  for i=1 to n do
    $res[i] := sim(q_k^T, H[i])$
  end for
  return $res$
```

The operation *retrieveClusterResult* as shown in Algorithm 6 returns a set of document clusters $Docs'$. Within this algorithm the central location of a cluster is computed via $(\frac{1}{|g_{rel}[j]=i|} \sum_{j \in \{1 \dots m\} \wedge g_{rel}[j]=i} S_j) \geq \theta$, where $|g_{rel}[j] = i|$ is the number of documents in the same group.

Algorithm 6 *retrieveClusterResult*

```
$retrieveClusterResult(Docs, g_{rel}, S, g, \theta)$:
  $Docs' := \emptyset$
  $m := size(Docs)$ //number of documents
  for i=1 to g do
    if $(\frac{1}{|g_{rel}[j]=i|} \sum_{j \in \{1 \dots m\} \wedge g_{rel}[j]=i} S_j) \geq \theta$ then
      $Docs' := Docs' \cup \bigcup_{j \in \{1 \dots m\} \wedge g_{rel}[j]=i} \{Docs[j]\}$
    end if
  end for
  return $Docs'$
```

The operation *retrieveClusterResult* delivers the information which documents are in the same cluster. All document representations of the same cluster are fused to a union A-box Sym'. If all input parameters for the repository $\mathcal{R}'$ are given, the query answering problems QA_{hol} and QA'_{sym} can be solved. The query answering problem $QA'_{\text{sym}}(.)$ is presented in the *SymQueryGroup* algorithm (see Algorithm 7). In comparison to the query answering problem QA_{sym}, QA'_{sym} computes document clusters and then a union A-box of clustered symbolic representations *Sym′* in order to solve the query answering problem QA'_{sym}.

Algorithm 7 The *SymQueryGroup* algorithm.

```
$QA'_{\text{sym}}(cq, (Docs', \_, \mathcal{T}, Sym, \_), \_)$:
  $docs' := \emptyset$
  for $group \in Docs'$ do
    $Sym' := \cup_{doc \in group} Sym_{doc}$
    $ans := answers((\mathcal{T}, Sym'), cq)$
    if $ans \neq \emptyset$ then
      $docs' := docs' \cup \{(group, ans)\}$
    end if
  end for
  return $docs'$
```

Example In an example we demonstrate IR results of QA_{hol} and of QA'_{sym}. Therefore we use our knowledge management scenario. The *clusterDocuments* operation computes the scores S and groups g_{rel} of the ten documents $d_1 \dots d_{10}$, the holistic representation H, and the string query "high jump" with the reduced query vector representation $\vec{q_2}$ (the computations of H and $\vec{q_2}$ are described in Subsection 4.1.2). The returned result is presented in Table 5.1.

Query answering results for QA_{hol} If the threshold $\theta = 0.94$, QA_{hol} will return:

$$docs = \{d_1, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

Document d_1 is a high jump document. The documents d_3, d_4, d_5, and d_6 are similar but do not especially represent high jump news. The documents d_7, d_8, and d_9 have no associations to high jump or athletics news. If the threshold is slightly higher, the result set *docs* for QA_{hol} will be:

$$docs = \{d_1, d_3, d_4, d_5, d_6, d_8, d_9\}.$$

The result set has less false positive documents: document d_7 is no longer in the result set.

$Docs$	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	1.00	1	3
d_2	0.92	2	4
d_3	1.00	1	4
d_4	1.00	1	4
d_5	1.00	1	4
d_6	1.00	1	4
d_7	0.94	2	2
d_8	1.00	1	1
d_9	1.00	1	2
d_{10}	0.01	3	2

Table 5.1: **Scores S of $Docs$ for the string query "*high jump*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H.**

Query answering results for $QA'_{\mathbf{sym}}$ The query answering QA'_{sym} has to solve $Docs'$ which is required first. $Docs'$ is delivered by *retrieveClusterResult*. If we call $retrieveClusterResult(Docs, g_{rel}, S, 3, 0.94)$ with $Docs = \langle d_1 \ldots d_{10} \rangle$, $g_{rel} = \langle 1, 2, 1, 1, 1, 1, 2, 1, 1, 3 \rangle$, and $S = \langle 1, 0.92, 1, 1, 1, 1, 0.94, 1, 1, 0.01 \rangle$, $Docs'$ will be:

$$Docs' = \{\{d_1, d_3, d_4, d_5, d_6, d_8, d_9\}, \{d_2, d_7\}, \{d_{10}\}\}.$$

Document d_2 is in a cluster although the score is smaller than the threshold. In the athletics context, d_2 is more similar to d_1 than to documents d_8 and d_9. QA'_{sym} returns the following documents:

$$docs = \{d_1, d_3, d_4, d_9, d_{10}\}.$$

The high jump document d_1 is in the result set, but there are also four false positive documents in the result set.

If the threshold is a bit higher ($\theta = 0.95$), $Docs'$ will be:

$$Docs' = \{\{d_1, d_3, d_4, d_5, d_6, d_8, d_9\}\}.$$

The number of returned cluster results is smaller, but the delivered results of QA'_{sym} is equal because QA'_{sym} returns the following documents:

$$docs = \{d_1, d_3, d_4, d_9, d_{10}\}.$$

If we increase the number of groups g to 4, *retrieveClusterResult*($Docs$, g_{rel}, S, 4, 0.94) with $Docs = \langle d_1 \dots d_{10} \rangle$, $g_{rel} = \langle 3, 4, 4, 4, 4, 4, 2, 1, 2, 2 \rangle$, and $S = \langle 1.00, 0.92, 1.00, 1.00, 1.00, 1.00, 0.94, 1.00, 1.00, 0.01 \rangle$ will deliver the result set $Docs'$:

$$Docs' = \{\{d_1\}, \{d_2, d_3, d_4, d_5, d_6\}, \{d_8\}\}.$$

The result set represents a correct clustering of the documents w.r.t. the string query "high jump." Nevertheless document d_8 does not represent high jump news. QA'_{sym} returns the following documents:

$$docs = \{d_1, d_2, d_3, d_4, d_5, d_6\}.$$

All delivered documents are about athletics. There are not documents about the fairy tale domain. However, the documents d_2, d_3, d_4, d_5, and d_6 are false positives. If we increase the threshold value to 0.95, the results will not change:

$$Docs' = \{\{d_1\}, \{d_2, d_3, d_4, d_5, d_6\}, \{d_8\}\}.$$

QA'_{sym} returns the following documents:

$$docs = \{d_1, d_2, d_3, d_4, d_5, d_6\}.$$

If the threshold is $\theta = 1.00$, $Docs'$ will be:

$$Docs' = \{\{d_1\}, \{d_8\}\}.$$

For the query "high jump" the document cluster is more precise. The documents d_2, d_3, d_4, d_5, and d_6 about athletics are not in the result set anymore. Document d_8 is in the result set, but it is not a document about athletics. QA'_{sym} returns the desired document:

$$docs = \{d_1\}.$$

There are not any false positive documents. However, in general, a threshold $\theta = 1.00$ is not a good choice. Next, we present how a semantics-based document reduction approach delivers more precise results.

Semantics-based Document Reduction

A variety of modeling methods including LSI have been proposed to solve IR tasks. These modeling methods are based on holistic approaches (i.e. PLSA). A combined holistic and symbolic approach is presented in the following. We show how to efficiently create the latent structure of documents using information from symbolic representations. The latent structure using symbolic representations is a complementarity matrix which represents the complementaries between all document representations. This new representation matrix is an input matrix of LSI. New document representations are computed. The computation of the document representation at lower rank based on symbolic representations is called semantics-based document reduction.

The semantics-based document reduction process is based on LSI and clustering as described above, whereby instead of the term-document matrix a so-called semantics-based complementarity matrix H_{Sem} is used and computed offline. H_{Sem} is defined as:

$$H_{Sem} := \begin{pmatrix} {h'}_{d_1,d_1} & {h'}_{d_2,d_1} & \dots & {h'}_{d_n,d_1} \\ {h'}_{d_1,d_2} & {h'}_{d_2,d_2} & \ddots & {h'}_{d_n,d_2} \\ \vdots & \ddots & \ddots & \vdots \\ {h'}_{d_1,d_n} & \dots & {h'}_{d_{n-1},d_n} & {h'}_{d_n,d_n} \end{pmatrix},$$

where

$${h'}_{d_i,d_j} := \left| Sym_{d_i} \setminus \Delta_{Sym_{d_i},Sym_{d_j}} \right| \text{ with } 1 \leq i,j \leq N.$$

In our example we receive the following complementarity matrix:

$$H_{Sem} = \left(\begin{array}{cccccc|cccc} 15 & 1 & 2 & 2 & 1 & 2 & 0 & 2 & 5 & 0 \\ 11 & 15 & 10 & 11 & 15 & 11 & 4 & 3 & 4 & 9 \\ 7 & 14 & 15 & 13 & 14 & 13 & 3 & 3 & 5 & 8 \\ 8 & 14 & 12 & 15 & 14 & 13 & 3 & 3 & 4 & 9 \\ 11 & 15 & 10 & 11 & 15 & 11 & 4 & 1 & 5 & 10 \\ 8 & 14 & 12 & 13 & 14 & 15 & 3 & 3 & 4 & 9 \\ \hline 15 & 14 & 10 & 11 & 11 & 14 & 15 & 5 & 6 & 4 \\ 1 & 12 & 8 & 9 & 15 & 9 & 15 & 15 & 7 & 7 \\ 12 & 14 & 11 & 12 & 13 & 11 & 8 & 0 & 14 & 5 \\ 13 & 12 & 12 & 12 & 12 & 12 & 5 & 3 & 7 & 14 \end{array} \right)$$

The diagonal values are the maximum because documents are compared to themselves. Large values represent low complementarity, and small values represent high complementarity. The example shows that the complementarity values of the six documents about athletics (above/left) and the four documents about fairy tales (below/right) are surprising high (below/left). And the complementarity values within the fairy tale are surprising low (below/right). Explanations for the high complementarity are that the fairy tale concepts are different and person names in fairy tales and athletics are similar. The computation of H_{Sem} can be done in an offline process.

The new holistic document representation H_{Sem} in a 2-dimensional space is computed via LSI, then:

$$V_2'^T = \begin{pmatrix} -0.31 & -0.42 & -0.34 & -0.37 & -0.42 & -0.37 & -0.20 & -0.12 & -0.20 & -0.26 \\ 0.61 & -0.04 & 0.07 & 0.05 & -0.16 & 0.03 & -0.54 & -0.55 & 0.03 & 0.05 \end{pmatrix}$$

The reduced column vectors of $V_2'^T$ represents the degree of associations between the documents. The reduced query vector $\vec{q_2}'$ for the string query "high jump" is computed in an online process: $q_2' = \langle -0.42, 0.51 \rangle$. If the new document representation H_{Sem} is used instead of H, the scores for each document will be a bit lower (cf. Table 5.1 and Table 5.2).

Docs	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	0.98	1	2
d_2	0.85	2	1
d_3	0.92	2	3
d_4	0.90	2	3
d_5	0.78	2	1
d_6	0.89	2	3
d_7	0.18	3	4
d_8	−0.05	3	4
d_9	0.91	2	3
d_{10}	0.91	2	3

Table 5.2: **Scores S of *Docs* for the string query "*high jump*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H_{Sem}.**

The scores in Table 5.2 show that d_1 and d_3 are more similar than the fairy tale documents d_8 and d_9. The documents d_7 and d_8 have a very low score. The scores of documents d_9 and d_{10} are very high because of the concept name Person. Athletics as well as fairy tale characters are persons so that the difference between the symbolic representations, i.e., Sym_{d_1} and Sym_{d_9}, are quite low. Therefore, the documents d_2, d_3, d_4, d_5, d_6, d_9, and d_{10} are in the same group (where $g = 3$), or d_3, d_4, d_6, d_9, and d_{10} are in the same group (where $g = 4$). The returned clusters $Docs'$ of *clusterDocuments* are presented which are input values for QA'_{sym}. In the following retrieval results of QA_{hol} and QA'_{sym} are presented.

Query answering results for $QA_{\mathbf{hol}}$ If the threshold is $\theta = 0.95$ and $g = 3$, the retrieval result of QA_{hol} will be:

$$docs = \{d_1\}.$$

The result set contains the correct document. There are no false positive or false negative documents in the result set. If the number of groups is $g = 4$, QA_{hol} will deliver the same result:

$$docs = \{d_1\}.$$

If the number of groups is $g = 4$, QA_{hol} will deliver the result set:

$$docs = \{d_1, d_3, d_4, d_9, d_{10}\}.$$

In the set are four false positives. The threshold $\theta = 0.9$ is not a good choice.

Query answering results for $QA'_{\mathbf{sym}}$ First, the document clusters are computed via *clusterDocuments*. The result set is:

$$Docs' = \{\{d_1\}, \{d_3, d_4, d_6, d_9, d_{10}\}\}.$$

Document d_6 is in a cluster because of the terms "high jump" found in the document. QA'_{sym} returns the following documents:

$$docs = \{d_1, d_3, d_4, d_6, d_9, d_{10}\}.$$

If the number of groups is reduced to $g = 3$, QA_{hol} with the additional usage of the *clusterDocuments* operation will deliver the result set:

$$Docs' = \{\{d_1\}\}.$$

QA'_{sym} returns the following document:

$$docs = \{d_1\}.$$

QA	$Hol.repr.$	θ	g	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
QA_{hol}	H	0.94	-	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
QA_{hol}	H	0.95	-	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗
QA'_{sym}	H_{sem}	0.94	3	✓	✗	✓	✓	✗	✗	✗	✗	✓	✓
QA'_{sym}	H_{sem}	0.95	3	✓	✗	✓	✓	✗	✗	✗	✗	✓	✓
QA'_{sym}	H_{sem}	0.94	4	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
QA'_{sym}	H_{sem}	0.95	4	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
QA'_{sym}	H_{sem}	1.00	4	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA_{hol}	H_{sem}	0.95	-	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA_{hol}	H_{sem}	0.90	-	✓	✗	✓	✓	✗	✗	✗	✗	✓	✓
QA'_{sym}	H_{sem}	0.90	3	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 5.3: **Retrieval results for QA_{hol} and QA'_{sym} for the string query "high jump." Hits are marked with ✓and no hits with ✗.**

The summarized results in Table 5.3 show that solving query answering problems of the form $QA_{\text{hol}}(\vec{q_k}', (Docs, H_{Sem}, \_, \_, \_), \theta)$ increase precision because of the new input parameter H_{Sem} whereby the threshold value has to be about 0.95. For smaller threshold values the additional use of the *clusterDocuments* operation increases the precision of IR result sets. More experiments and discussions are given in Section 5.2.

Semantics-based Information Retrieval

In the case of retrieval results with high recall ($|docs'| \geq \delta$), it should be possible to translate holistic queries into symbolic ones automatically, so that a user must not be an expert in query languages. An approach for defining particular translations from string queries to symbolic ones is presented in [EKS13]. Here, the according algorithm is called *HolSym* and is presented in Algorithm 8. In Algorithm 8 a query vector $\vec{q_k}$ is an input parameter.

For computing $docs''$ the query is transformed into a symbolic query with the operation $transform2cq(\vec{q_k})$. For implementing the transformation the approach presented in [EKS13] is recommended. The transformation function $transform2cv(docs, H_{sem}) := |docs|^{-1} \sum_{doc_i \in docs}(H_{sem}[i])$ computes a necessary complementary vector $\vec{cv}$. In Algorithm 8 the *transform2cv* is used in order to receive document indicators for further computations.

For retrieval results with very low recall ($|docs'| \leq \delta$), is should also be possible to translate symbolic queries into holistic ones automatically. The according algorithm is called *SymHol* and is presented in Algorithm 9. In Algorithm 9 a symbolic query *cq* is an input parameter. For computing $docs''$ the query is transformed into a string query with the operation $transform2qk(cq)$.

Algorithm 8 *HolSymMethodology.*

```
$HolSym(\vec{q_k}, ((Docs, H, \mathcal{T}, Sym, \_), H_{sem}), \theta, g, \delta)$:
  $docs' := QA_{\mathrm{hol}}(\vec{q_k}, (Docs, H, \_, \_, \_), \theta)$
  if $|docs'| \geq \delta$ then
    $(g_{rel}, S) := clusterDocuments(transform2cv(docs', H_{sem}), H_{sem}, g)$
    $Docs' := retrieveClusterResult(Docs, g_{rel}, S, g, \theta)$
    $docs'' := QA'_{\mathrm{sym}}(transform2cq(\vec{q_k}), (Docs', \_, \mathcal{T}, Sym, \_), \_)$
  else
    $docs'' := \emptyset$
  end if
  return $(docs', docs'')$
```

Algorithm 9 *SymHolMethodology.*

```
$SymHol(cq, ((Docs, H, \mathcal{T}, Sym, \_), H_{sem}), \theta, g, \delta)$:
  $docs := QA_{\mathrm{sym}}(cq, (Docs, \_, \mathcal{T}, Sym, \_), \_)$
  $(g_{rel}, S) := clusterDocuments(transform2cv(docs, H_{sem}), H_{sem}, g)$
  $Docs' := retrieveClusterResult(Docs, g_{rel}, S, g, \theta)$
  $docs' := QA'_{\mathrm{sym}}(cq, (Docs', \_, \mathcal{T}, Sym, \_), \_)$
  if $|docs'| \leq \delta$ then
    $docs'' := QA_{\mathrm{hol}}(transform2q_k(cq), (Docs, H, \_, \_, \_), \theta)$
  else
    $docs'' := \emptyset$
  end if
  return $(docs', docs'')$
```

As a result the Algorithms 8 and 9 return a tuple of documents $(docs', docs'')$.

Quality and performance issues of the *HolSym Methodology* are discussed in the next section.

In the web exist documents which cannot be delivered by IR systems because publishers do not use proper search terms for their texts. In order to capture such documents, we suggest to use logistic tensor factorization for multi-relational data as a suitable approach for computing explicit knowledge by predicting latent structures of missed explicit terms for a true positive search result. A tensor factorization method is i.e. Rescal. It is used for efficiently modeling, analyzing, and predicting data with multiple modalities. Dyadic relational data with D different relations and N entities, has a natural representation as an adjacency tensor X of size $N \times N \times K$, where $x_{ijk} = 1$, if the $\mathit{Relation}_k(\text{Entity}_i, \text{Entity}_j)$ exists, and 0 otherwise. Rescal is a latent factor model for relational learning. Rescal factorizes an adjacency tensor X into latent representations of entities and relations. Formally:

$$X_k \approx A R_k A^T, \tag{5.3}$$

where X_k is the k-th factor matrix of X, matrix $A \in \mathbb{R}^{N \times r}$ holds the latent representations for the entities, $R_k \in \mathbb{R}^{r \times r}$ is the latent representation of the k-th predicate, and r is referred as the rank. Equation 5.3 can be also written as $x_{ijk} \approx a_i^T R_k a_j$, where the column vector $a_i \in \mathbb{R}^r$ denotes the i-th row of A. Rescal is applicable if latent factors are suitable for capturing essential information in a domain [NT13].

Another approach for enriching documents with explicit knowledge is presented in [BKM]. This contribution presents the approach *unsupervised text annotation* (UTA). UTA can be used to link documents with an associated symbolic content descriptions. The authors plan to extend UTA to an approach with an arbitrary query as input so that IR systems can return a set of documents answering the query.

5.2 Evaluation

This chapter is structured as follows: First, we explore the feasibility of the A-box difference operator and LSI through a representative experimental study. Second, we analyze the quality of the A-box fusion algorithm (Algorithm 1) compared to the fusion approach by inference (cf. [Kay11]). For the evaluation of the quality we exploit a large corpus of documents, which have been annotated by human experts using concept and role names from respective domain ontologies. Last but not least we present quality results for the suggested methodology for combining holistic and symbolic approaches for information retrieval.

5.2.1 Feasibility

In this section we analyze the performance and quality of our methodology through a representative experimental study in the following way:

- Start with a corpus of documents
- Collect a set of queries for this corpus
- Create the gold standard
- Measure the retrieval results
- Evaluate the retrieval results

We use a test corpus consisting of about 600 documents taken from the athletics domain (BOEMIE repository) for evaluating the A-box difference operation and 120 documents for the combination approach. The fusion experiments were run on a Windows 7 Enterprise 64-bit system with an Intel(R) Core^{TM} i5 3.2 GHz processor and 6 GB of main memory. The combination experiments were run on a Windows^{TM} system with an Intel(R) Core^{TM}2 Duo CPU, 2 GHz processor and 2 GB of main memory. The algorithms were implemented in Eclipse version 4.2.1. The DL reasoner RacerPro version 2.0 was used as an extern program in order to execute the A-box fusion algorithm.

Feasibility of A-box Difference Computation

Table 5.4 and Figure 5.1 indicate the performance of the A-box difference operator.

Table 5.4 shows that the time of an A-box difference operation of two A-boxes with an average size of 1.5 kB is reasonable. With an A-box size of 13 kB the computation time increases drastically (see Figure 5.1).

KBytes	Milliseconds	KBytes	Milliseconds
3.16113	4213	8.26270	3445
3.17383	7404	9.41406	7798
3.83398	8435	9.86816	17058
3.94531	7734	9.91797	64937
6.16016	16940	10.1162	41279
6.34375	4702	10.5606	13619
6.64746	3891	11.4297	114236
7.27051	3640	12.8145	570820
7.56445	19605	13.6152	3436459
7.71777	8748	15.3125	4553505

Table 5.4: A-box sizes and performance of A-box difference operations

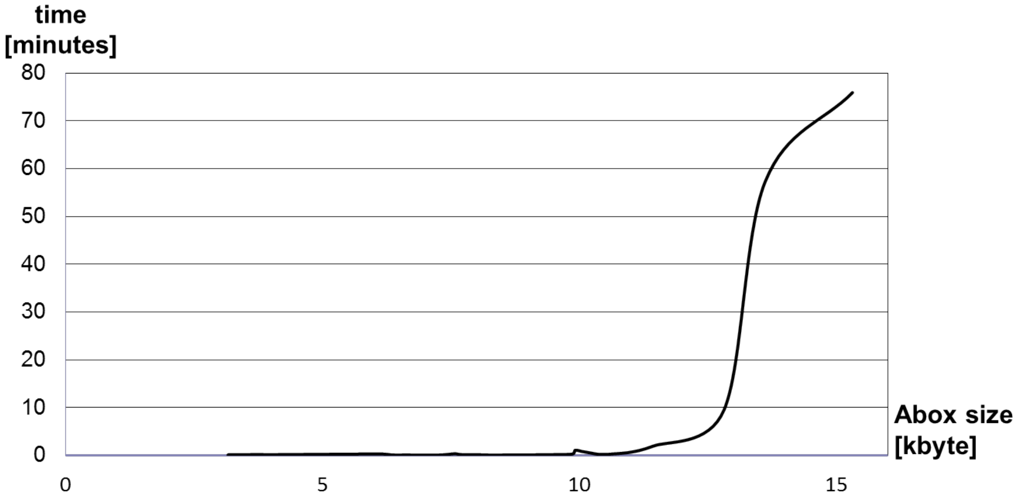


Figure 5.1: Performance of A-box difference operations

This experiment shows that performance is quite low for an A-box difference operation with an A-box size upwards to 13 kB. The performance results are acceptable for small A-box sizes but hardly practical for larger ones. In [Wan11] an approach is presented how to handle with growing sets of assertional statements in ontologies. The aim of this work is to reduce instance checking for an individual in an ontology to smaller subsets. To this end better performance results should be achievable.

Feasibility of LSI

In general, the performance results of LSI or more concretely SVD depends on the number of k and the number of terms and documents. Some experiments in the literature indicate that a value of $k = 300$ to 500 provides best performance results. For a five million document collection, a value of k $\approx$ 400 provides the best performance [Bra08].

In our experiment we have used 41, 1682, and 45687 documents of the BOEMIE repository and the ten terms *HighJump*, *PoleVault*, *Athlete*, *Person*, *Marathon*, *SportsTrial*, *Performance*, *City*, *PersonName*, and *Rank* for indexing as representative terms for demonstrating the performance results. In Figure 5.2 it is presented that the performance decreases linearly with an increasing number of documents in the repository. For instance, the computation time for a 10×41 matrix is about 2 milliseconds and for a 10×1682 matrix is about 1 second (1427 milliseconds), and for a 10×45687 matrix is about 25 minutes (1514592 milliseconds). The optimal choice of k depends on the quality of terms which are used. In Subsection 5.2.2 we discuss the best choice of k.

Similarity Score of the Topic Result Set

LSI provides a basis in the field of topic modeling. In order to define a result set with relevant documents on the basis of a topic result set (document $V_k^T[i]$ vs. query $\vec{q_k}$) a score value $score_i$ is required. For example (approximate) nearest neighbor algorithms such as *Cosine Similarity* or *Locality Sensitive Hashing* (LSH) delivers the required scores. The nearest neighbor problem is the following: Given a set V_k^T of i points in a metric space defined over a set

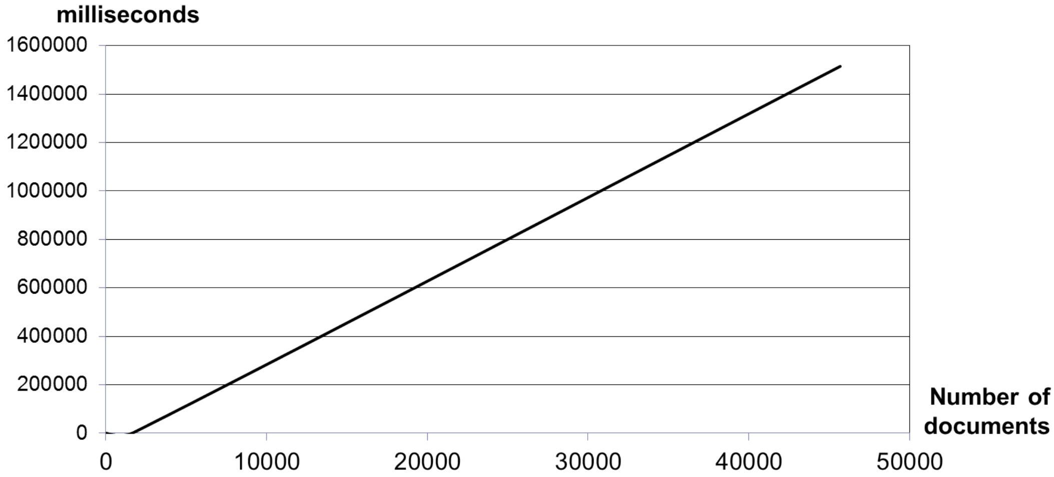


Figure 5.2: Performance of LSI

X with a distance function D, pre-process V_k^T to efficiently answer queries $\vec{q_k}$ for finding the points in $V_k^T[i]$ closest to a query $\vec{q_k} \in X$, where $X = \mathbb{R}^d$ with dimension d under some l_s norm [IM98].

Cosine Similarity Cosine similarity measure is used in information retrieval system if it is important to have a low false positive rate. Relevant documents on the basis of topic result set (document $V_k^T[i]$ vs. query $\vec{q_k}$) are computed. If the score *score* of a document vs. the query is within a given threshold value Θ, the document will be returned to the user. Unfortunately, as the dimension increases, this approach become less efficient because the space or time requirements increase exponentially in the dimension [IM98]. We suggest to use locality-sensitive hashing in Algorithm 2 in order to achieve better performance results.

5.2.2 IR Quality

In this section, we compare and evaluate the quality of precision of information results using the fusion approaches presented in Section 4.2.1: fusion using a) the A-box difference operator and b) structure-oriented rules.

Quality of Documents using the A-box Difference Operator

In Section 4.2.1 we have presented that the abduction process computes two different explanations for an high jump image (cf. Figure 4.15 and Figure 4.16). Namely, high jump and pole vault. In the case, if more than one explanation exists for an image, we suggest to use the a-box difference operator in order to increase precision of document representations. Therefore the knowledge of basis and differences of interpretation a-boxes is required. In [MGK$^+$14] we present how to compute the basis and the *semantic* differences of two a-boxes. In addition we present that a T-box also plays an important role during the computation of differences. If the T-box has the CGI $\mathit{HighJump} \sqcap \mathit{PoleVault} \sqsubseteq \bot$ (high jump and pole vault are disjoint) then it is a hint that one explanation is only valid. For making a decision for one symbolic representation further knowledge is required. We suggest to use an associated A-box caption or text and fuse the A-boxes (image and caption/ image and text). If the fused A-box is consistent then we can decide for the more precise symbolic representation. In our example, we can decide for the high jump representation (Figure 4.15). A-box difference operator increases the precision of interpretation A-boxes.

In some cases in an A-box there are different individual names which represent the same object (see Figure 5.3). In Figure 5.3 the number of individuals of 78 A-boxes is presented. We see that the number of same-as individuals increases with the number of total individuals.

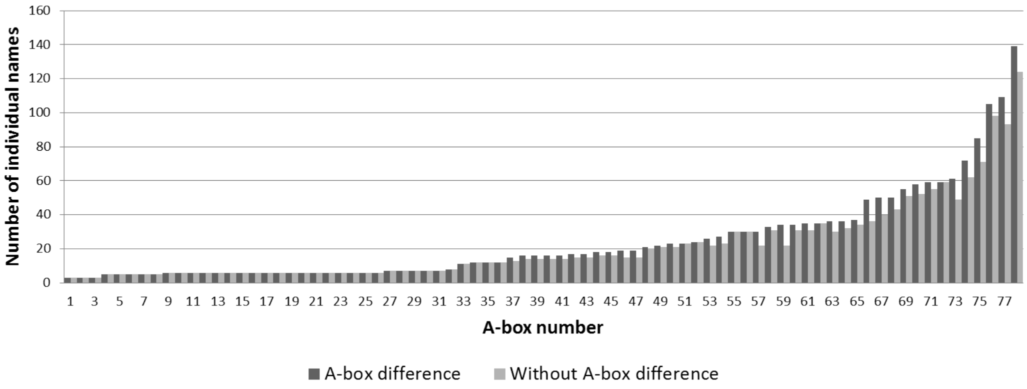


Figure 5.3: The number of individual names with and without using the A-box difference operator

In order to evaluate IR results, in our experiment we compute the difference between image and caption A-Boxes for 78 documents using the A-box difference operator and create for each document a fusion A-box so that the new same-as assertions are parts of a fusion A-box. In sum, the number of retrieved individuals is 1630 using the A-box difference operator and 1432 without using a fusion algorithm. All retrieved individual names are relevant in our experiment, and the false positive rate decreases by 13.82 percent (cf. [Cau13]).

A-box Fusion Algorithm versus Fusion with Structure-Oriented Rules

In a further experiment we compare A-box fusion algorithm with the fusion approach using structure-oriented rules. Again, we use a repository with 78 documents and their symbolic representations (A-boxes) and compute the number of different individuals. Figure 5.4 shows that the number of same-as individuals could be reduced using the A-box difference operator compared with the fusion approach with structure-oriented rules by the factor of 91 percent.

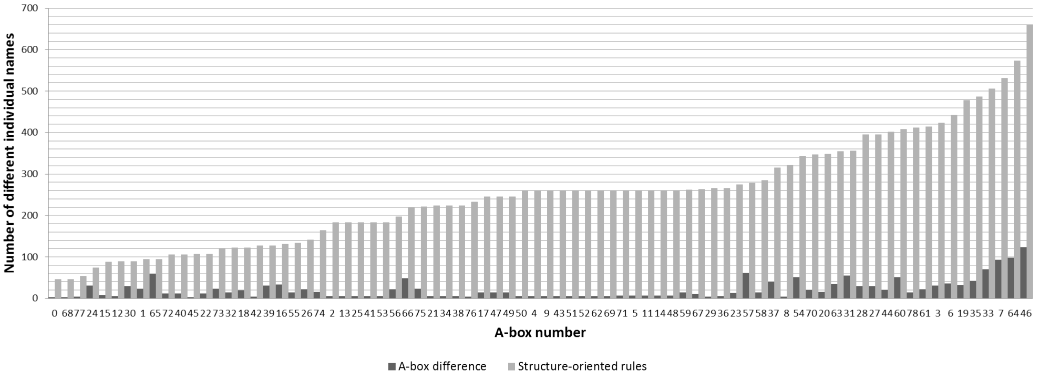


Figure 5.4: The number of different individuals in an A-Box using the A-box difference operator and the fusion approach using structure-oriented rules

The reduction of equal individuals has the advantage that the A-box sizes are smaller, so that the A-box difference supports a better performance than the fusion approach using structure-oriented rules.

Quality of LSI Results

We study LSI research results in order to evaluate the quality of LSI. Some studies of LSI present that for a low k the performance is quite well, but the quality of IR results decreases because the representation of relationship between document and terms is not representative [Bra08]. Hence, the choice of the parameter k is essential w.r.t. performance and quality of IR results.

In Section 4.1 we presented that the best approximation is $k = 4$ in our knowledge management scenario for a small repository. For a repository with up to 5 million documents an optimal k is between 300 and 500 for specific tasks (cf. Table 1 in [Bra08]). In general there is a great variability w.r.t. the kind of documents, the used terms and queries. For instance, the studies in Table 5.5 use encyclopedias, articles, emails, or documents in different languages etc. as an input repository. An optimal k of these studies has been found to be between 80 and 400 for specific tasks.

k	Mean (#docs)	Median(#docs)	References
400	10377	1238	[JM03, LG03, LST07]
300	17613	5939	[LD97, WHG99, JL00, AF02, AKS06, BRP07]
250	37600	37600	[KKH00, TL03]
200	12649	2146	[Hul94, SLD96, YCBF98, WH00, HSD01, Gee03b, KCZ03, KP06]
150	699	699	[ZMS98, Pin04]
100	3917	1217	[DDF$^+$90a, LS01, Che03, Mor05, YYT05, Gei06]
90	1803	1803	[Dum03], in [Bra08] ref. [36]
80	2774	1000	[MBW05, HTDRP07]

Table 5.5: **Evaluations results for an optimal LSI dimensionality.**

In [Bra08] Bradford discusses that LSI studies have dealt with small test collections, so that chance co-occurrence of terms in few documents can have significant impact on term-term relationship. In order to find out the optimal k Bradford uses a large test collection. As a result he receives $k = 400$ as a best choice.

In order to find a mathematical explanation for the choice of k for a repository, we have compared all results of the LSI studies so that different kind of documents were considered. We determine for each k the mean and median of test set size of documents (#docs). The values are presented in Table 5.5, and a graphically representation is given in Figure 5.5.

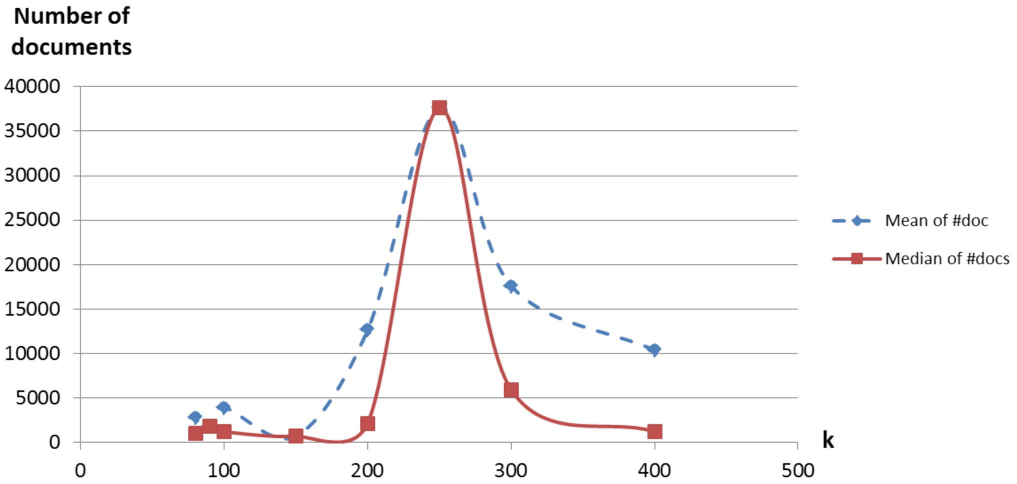


Figure 5.5: Plot of measured optimum parameter k between 80 and 400 w.r.t. the mean and median of test set size of documents (#docs).

Figure 5.5 presents interesting information:

- The optimal parameter k is 250.
- The median curve represents a normal distribution.

The formula for the normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{\frac{-(x-\mu)^2}{2\sigma^2}}, \tag{5.4}$$

where μ is the mean and σ the standard deviation. Our opinion is that the parameter μ represents the value k. Standard normal distribution values are $\mu = 0$ and $\sigma^2 = 1$. In our example $\mu = 250$ and $\sigma = \pm 150$.

In the research field of (automatic) text processing it is known that significant terms follows a normal distribution placed over a term list ranked by the frequency of term occurrence [LH11]. In this work we establish the link between the optimal parameter k, an input parameter for LSI and the normal distribution.

Quality of the HolSym Methodology

The *clusterDocuments* operation is one essential operator of the HolSym Methodology. We evaluate this operator using our knowledge management scenario. We compute scores S and groups g_{rel} of the ten documents $d_1 \dots d_{10}$, use the respective holistic representation H and the string queries “pole vault,” ”Prince Charming,” and ”Snow White” with the particular reduced query vector representation $\vec{q_2}$ (the computations of H and $\vec{q_2}$ are described in Subsection 4.1.2). The returned results are presented

- in Table 5.6 and Table 5.7 for the query “pole vault.”
- in Table 5.9 and Table 5.10 for the query “Prince Charming.”
- in Table 5.12 and Table 5.13 for the query “Snow White.”

In the following we compare the result sets between QA_{hol} and QA'_{sym}.

IR Example: “Pole Vault” For the query “pole vault” and a threshold $\theta = 0.95$, QA_{hol} delivers (cf. Table 5.6):

$$docs = \{d_1, d_3, d_4, d_5, d_6, d_8, d_9\}.$$

Document d_2 is a pole vault document and the result set does not contain d_2. The documents d_1, d_3, d_4, d_5, and d_6 are similar but do not especially represent pole vault news. The documents d_8 and d_9 have no associations to pole vault or athletics news. If the threshold is 0.92, the result set *docs* for QA_{hol} will be:

$$docs = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

The result set has more false positive documents, but d_2 is in the result set.

For solving QA_{sem} it is required to compute document clusters first. If the $retrieveClusterResult(Docs, g_{rel}, S, 3, 0.95)$ with $Docs = \langle d_1, d_2, d_3, d_4, \dots, d_{10}\rangle$, the score values $S = \langle 0.84, 0.98, 1.00, 1.00, 0.95, 0.99, 0.52, 0.30, 1.00, 1.00\rangle$, and $g_{rel} = \langle 1, 2, 2, 2, 2, 2, 3, 3, 2, 2\rangle$, $Docs'$ will be (cf. Table 5.7):

$$Docs' = \{\{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}\}.$$

The returned documents fo QA_{sem} are:

$$docs = \{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}.$$

Document d_2 is in the result set, but there also are false positive documents. And d_5 is a false negative document.

If the number of groups g is 4, *retrieveClusterResult* ($Docs$, g_{rel}, S, 4, 0.95) with $Docs = \langle d_1 \ldots d_{10} \rangle$, $g_{rel} = \langle 2, 1, 3, 3, 1, 3, 4, 4, 3, 3 \rangle$, and the score values $S = \langle 0.84, 0.98, 1.00, 1.00, 0.95, 0.99, 0.52, 0.30, 1.00, 1.00 \rangle$ will deliver the following result set $Docs'$ (cf. Table 5.7):

$$Docs' = \{\{d_2, d_5\}, \{d_3, d_4, d_6, d_9, d_{10}\}\}.$$

The result set represents a good clustering of the documents w.r.t. the string query "pole vault." The returned documents fo QA_{sem} are $docs = \{d_2, d_5\}$. The returned documents d_2 and d_5 correctly represents "pault vault" news. This example shows that better cluster results could be expected, if the matrix H_{sym} has a better semantics-based representation of documents. A summarized overview about the IR results is given in Table 5.8.

$Docs$	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	1.00	1	1
d_2	0.92	1	1
d_3	1.00	1	1
d_4	1.00	1	1
d_5	1.00	1	1
d_6	1.00	1	1
d_7	0.94	1	4
d_8	1.00	1	4
d_9	1.00	2	2
d_{10}	0.01	3	3

Table 5.6: **Scores S of $Docs$ for the string query "*pole vault*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H.**

In the case where the precision of retrieved results is high, the operation $HolSym(\vec{q_k}, (Docs, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, Sym', \_), 0.95, 4, 3)$ is suitable to use for increasing precision. The returned results are:

$$(docs', docs'') = (\{d_1, d_3, d_4, d_5, d_6, d_8, d_9\}, \{d_2, d_5\})$$

$Docs$	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	0.84	1	2
d_2	0.98	2	1
d_3	1.00	2	3
d_4	1.00	2	3
d_5	0.95	2	1
d_6	0.99	2	3
d_7	0.52	3	4
d_8	0.30	3	4
d_9	1.00	2	3
d_{10}	1.00	2	3

Table 5.7: **Scores S of $Docs$ for the string query "*pole vault*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H_{sym}.**

QA	$Hol.repr.$	θ	g	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
QA_{hol}	H	0.95	-	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗
QA_{hol}	H	0.92	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
QA'_{sym}	H_{sem}	0.95	3	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓
QA'_{sym}	H_{sem}	0.95	4	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗

Table 5.8: **Retrieval results for QA_{hol} and QA'_{sym} for the string query "pole vault."** Hits are marked with ✓and no hits with ✗.

In the case where retrieved results are low by using symbolic query answering, the operation *SymHol* increases the recall:

$$(docs', docs'') = (\{d_2, d_5\}, \{d_1, d_3, d_4, d_5, d_6, d_8, d_9\})$$

In our example it makes sense so rank the results by precision and score *HolSym* should return $docs'' \cup docs'$, and *SymHol* should return $docs' \cup docs''$. The user would retrieve the documents d_2, d_5, d_3, d_4, d_9, d_6, d_1, and d_8. As a result precision and recall is increased.

IR Example: "Prince Charming" For the query "Prince Charming" and a threshold $\theta = 0.95$, QA_{hol} delivers (cf. Table 5.9):

$$docs = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

This result set contains many false positive documents. Only document d_7 is about Price Charming. If the threshold is 0.98, the result set *docs* for QA_{hol} will be:

$$docs = \{d_1, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

The result set has less false positive documents, and document d_7 is in the result set. If the threshold is 0.99, QA_{hol} will return two documents:

$$docs = \{d_8, d_9\}.$$

The result set has less false positive documents, but document d_7 is not anymore in the result set.

Before we compute QA'_{sym} for doing comparisons between retrieved IR results, semantics-based clustering is done. The returned clusters of *retrieve−ClusterResult*$(Docs, g_{rel}, S, 3, 0.95)$ with $Docs = \langle d_1 \dots d_{10} \rangle$, the score values $S = \langle 0.87, 0.97, 0.99, 0.99, 0.93, 0.98, 0.46, 0.24, 0.99, 0.99 \rangle$, and the group numbers $g_{rel} = \langle 1, 2, 2, 2, 2, 2, 3, 3, 2, 2 \rangle$ are (cf. Table 5.10):

$$Docs' = \{\{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}, \{d_7, d_8\}\}.$$

$Docs'$ has two clusters. The first cluster with documents $d_2, d_3, d_4, d_5, d_6, d_9, d_{10}$ only has false positive documents. The second cluster with documents $\{d_7, d_8\}$ has one right document (d_7) and one false positive document (d_8), whereby d_7 and d_8 are about the same domain. QA'_{sym} returns:

$$docs = \{d_7, d_8\}.$$

If the threshold is 0.98 or 0.99, $Docs'$ will be (cf. Table 5.10):

$$Docs' = \{\{d_7, d_8\}\}.$$

$Docs'$ has one cluster. In the cluster is one right (d_7) and one false positive document (d_8). QA'_{sym} returns the same documents:

$$docs = \{d_7, d_8\}.$$

If the number of groups g is increased to 4, *retrieveClusterResult*$(Docs, g_{rel}, S, 4, 0.95)$ with $Docs = \langle d_1 \dots d_{10} \rangle$, $g_{rel} = \langle 2,1,3,3,1,3,4,4,3,3 \rangle$, and $S = \langle 0.87, 0.97, 0.99, 0.99, 0.93, 0.98, 0.46, 0.24, 0.99, 0.99 \rangle$ will deliver the result set $Docs'$ (cf. Table 5.10):

$$Docs' = \{\{d_5\}, \{d_3, d_4, d_6, d_9, d_{10}\}\}.$$

Document d_7 is not part of a cluster. However, QA'_{sym} returns an empty set:

$$docs = \{\}.$$

If the threshold is 0.98 or 0.99, then $Docs' = \{\{d_3, d_4, d_6, d_9, d_{10}\}\}$. QA'_{sym} also returns an empty set $docs = \{\}$.

This example shows that the choice of $g = 4$ is not a good choice. A summarized overview about the IR results is given in Table 5.11.

$Docs$	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	0.98	1	1
d_2	0.97	1	1
d_3	0.98	1	1
d_4	0.98	1	1
d_5	0.98	1	1
d_6	0.98	1	1
d_7	0.98	1	4
d_8	0.99	1	4
d_9	0.99	2	2
d_{10}	0.18	3	3

Table 5.9: **Scores S of $Docs$ for the string query "*prince charming*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H.**

$Docs$	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	0.87	1	2
d_2	0.97	2	1
d_3	0.99	2	3
d_4	0.99	2	3
d_5	0.93	2	1
d_6	0.98	2	3
d_7	0.46	3	4
d_8	0.24	3	4
d_9	0.99	2	3
d_{10}	0.99	2	3

Table 5.10: **Scores S of $Docs$ for the string query "*prince charming*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H_{sym}.**

QA	$Hol.repr.$	θ	g	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
QA_{hol}	H	0.95	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
QA_{hol}	H	0.98	-	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗
QA_{hol}	H	0.99	-	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
QA'_{sym}	H_{sem}	0.95	3	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
QA'_{sym}	H_{sem}	0.98	3	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
QA'_{sym}	H_{sem}	0.99	3	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
QA'_{sym}	H_{sem}	0.95	4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA'_{sym}	H_{sem}	0.98	4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA'_{sym}	H_{sem}	0.99	4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Table 5.11: **Retrieval results for QA_{hol} and QA'_{sym} for the string query "Prince Charming."** Hits are marked with ✓and no hits with ✗.

In the case where the precision of retrieved results is high, the operation $HolSym(\vec{q_k}, (Docs, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, Sym', \_), 0.98, 4, 3)$ is suitable to use for increasing precision. The returned results are:

$$(docs', docs'') = (\{d_1, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}, \{d_7, d_8\})$$

In the case where retrieved results are low by using symbolic query answering, the operation $\mathit{SymHol}(cq, (\mathit{Docs}, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, \mathit{Sym}', \_), 0.95, 3, 3)$ increases the recall:

$$(\mathit{docs}', \mathit{docs}'') = (\{d_7, d_8\}, \{d_1, d_3, d_4, d_5, d_6, d_7, d_8, d_9\})$$

In our example it makes sense so rank the results by precision and score *HolSym* should return $\mathit{docs}'' \cup \mathit{docs}'$, and *SymHol* should return $\mathit{docs}' \cup \mathit{docs}''$. The user would retrieve the documents d_7, d_8, d_3, d_9, d_4, d_6, d_5, and d_1. In the second case the recall is increased.

IR Example: "Snow White" For the query "snow white" and a threshold $\theta = 0.95$, QA_{hol} delivers (cf. Table 5.12):

$$\mathit{docs} = \{d_{10}\}.$$

Document d_{10} is a snow white document. There are no false positive documents. Before we compute QA'_{sym} for doing comparisons between retrieved IR results, semantics-based clustering is done. The *retrieveClusterResult*(*Docs*, g_{rel}, S, 3, 0.95) with $\mathit{Docs} = \langle d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10} \rangle$, the scoring values $S = \langle 0.26, 0.86, 0.78, 0.79, 0.92, 0.81, 0.96, 0.86, 0.79, 0.78 \rangle$, and $g_{rel} = \langle 1, 2, 2, 2, 2, 2, 3, 3, 2, 2 \rangle$, the new result set Docs' will be (cf. Table 5.13):

$$\mathit{Docs}' = \{\{\}\}.$$

QA'_{sym} delivers no documents because $\mathit{docs} = \{\}$.

If the number of groups g is increased to 4, *retrieveClusterResult*(*Docs*, g_{rel}, S, 4, 0.95) with $\mathit{Docs} = \langle d_1 \ldots d_{10} \rangle$, $g_{rel} = \langle 2, 1, 3, 3, 1, 3, 4, 4, 3, 3 \rangle$, and $S = \langle 0.26, 0.86, 0.78, 0.79, 0.92, 0.81, 0.96, 0.86, 0.79, 0.78 \rangle$ will deliver the result set Docs' (cf. Table 5.13):

$$\mathit{Docs}' = \{\{\}\}.$$

QA'_{sym} delivers $\mathit{docs} = \{\}$ The result set is empty and represents that H_{sym} has not reached his optimal document representation. A summarized overview about the IR results is given in Table 5.14.

Docs	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	-0.04	1	1
d_2	0.39	1	1
d_3	-0.03	1	1
d_4	-0.04	1	1
d_5	-0.04	1	1
d_6	-0.03	1	1
d_7	0.34	1	4
d_8	0.06	1	4
d_9	0.00	2	2
d_{10}	1.00	3	3

Table 5.12: **Scores S of *Docs* for the string query "*snow white*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H.**

Docs	S	g_{rel} (with $g = 3$)	g_{rel} (with $g = 4$)
d_1	0.26	1	2
d_2	0.86	2	1
d_3	0.78	2	3
d_4	0.79	2	3
d_5	0.92	2	1
d_6	0.81	2	3
d_7	0.96	3	4
d_8	0.86	3	4
d_9	0.79	2	3
d_{10}	0.78	2	3

Table 5.13: **Scores S of *Docs* for the string query "*snow white*" and g_{rel} with $g = 3$ and $g = 4$ (right) with input matrix H_{sym}.**

QA	$Hol.repr.$	θ	g	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
QA_{hol}	H	0.95	-	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
QA'_{sym}	H_{sem}	0.95	3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA'_{sym}	H_{sem}	0.95	4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
QA'_{sym}	H_{sem}	0.99	3	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗

Table 5.14: **Retrieval results for QA_{hol} and QA'_{sym} for the string query "Snow White."** Hits are marked with ✓and no hits with ✗.

The operation $HolSym(\vec{q_k}, (Docs, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, Sym', \_), 0.95, 3, 3)$ returns in this example:

$$(docs', docs'') = (\{d_{10}\}, \{\})$$

In this case recall can be increased if the threshold is lower than 0.96, whereby the precision would be decreased.

5.3 Formalized Knowledge-Creation Process

Nonaka et al. define a knowledge creation process for creating knowledge in companies (see Subsection 2.2.1). The process of knowledge creation is based on the SECI process, a platform for knowledge creation, and so-called knowledge assets. The implementation of the knowledge creation process is called ba. In Subsection 2.2.1 we have discussed that the knowledge creation process as defined by Nonaka et al. is too informal because, in general, there only is a vague explanation "HOW" to create and share knowledge and it is not specified "WITH WHAT" the creation is accomplished.

In this subsection we present that Nonaka's visionary model is practically realizable by (i) defining operations for the SECI model (ii) using Nonaka's method for knowledge creation as a basis, and (iii) using semantic assets as introduced above as an implementation of Nonaka's knowledge assets.

The concrete symbolic knowledge creation process is illustrated in Figure 5.6. It represents a formalized symbolic knowledge creation process (SKCP), i.e., a SKCP with concrete operations. The SKCP has the following elements: SECI, *ba*, and semantic assets; and the operations: *create low-level content*

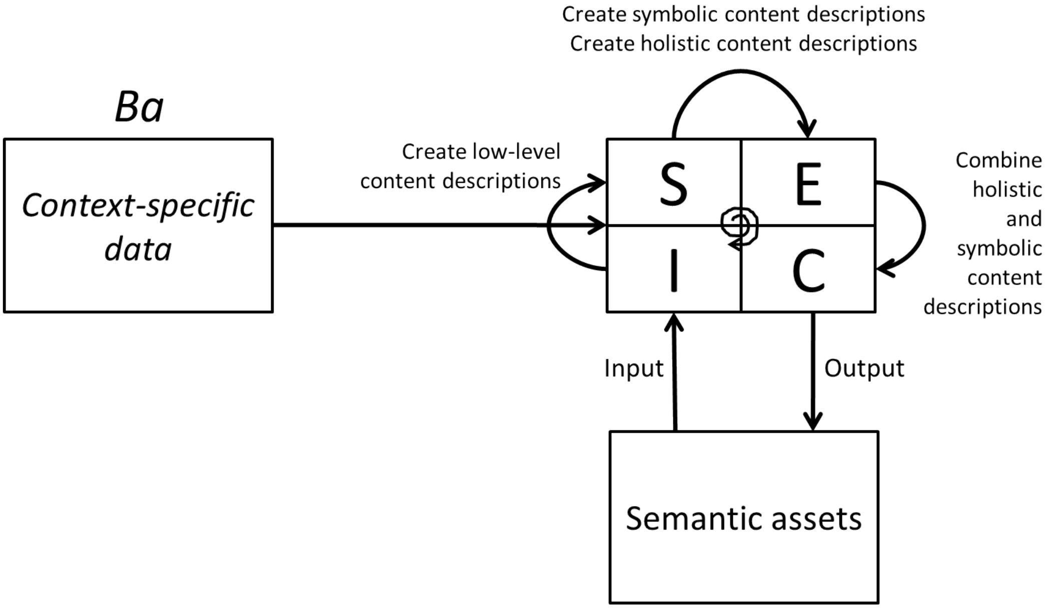

Create symbolic content descriptions
Create holistic content descriptions
Ba
Context-specific data
Create low-level content descriptions
S
E
I
C
Combine holistic and symbolic content descriptions
Input
Output
Semantic assets

Figure 5.6: Formalized symbolic knowledge creation process with the three elements: SECI, *ba*, and semantic assets; and the operations: *create low-level content descriptions*, *create symbolic content descriptions*, *create holistic content descriptions*, and *combine holistic and symbolic content descriptions.*

descriptions, *create symbolic content descriptions*, *create holistic content descriptions*, and *combine holistic and symbolic content descriptions.*

The operations of the SECI process transfer knowledge to the modes socialization, externalization, combination, and internalization. In the following, we describe the operations and new assets types, and we present an example afterwards.

Create low-level content descriptions The operator *create low-level content descriptions* generates a transfer from the mode initialization to socialization. The transfer is done via an analysis process.

For the internalization process there exists a repository (e.g., derived from a set of web pages) without any content descriptions. The content of documents, or of parts of documents (i.e. images, texts) in a repository is represented by an asset which we call pre-iconographical asset (pre-asset).

For each document or document parts low-level content descriptions will be created. These low-level descriptions are called iconographical asset (ico-asset).

Create symbolic and holistic content descriptions The operators *create symbolic content descriptions* and *create holistic content descriptions* generate a transfer from the mode socialization to externalization. The transfer is done via an interpretation process.

For the socialization process there exists ico-assets. The operators *create symbolic content descriptions* and *create holistic content descriptions* create high-level content descriptions. Both processes are described in detail in Sections 3.2, and 3.3, respectively. The high-level content descriptions are represented by an asset which we call iconological asset (log-asset).

Combine holistic and symbolic content descriptions The operator *combine holistic and symbolic content descriptions* generates a transfer from mode externalization to combination. The transfer is done via the HolSym Methodology.

The outputs of the HolSym Methodology are called combined asset (com-asset).

Semantic assets *Semantic assets* are the input and output assets of the SECI process. We presented four different types of semantic assets: pre-asset, ico-asset, log-asset, and com-asset.

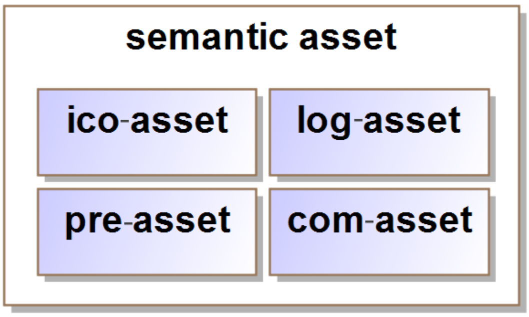


Figure 5.7: **Semantic asset types:** pre-asset, ico-asset, log-asset, and com-asset.

In the following we present an example showing how semantic assets are created.

SECI process In general the SECI process creates semantic assets for a multimedia document at different levels (see Figure 5.8). At level 0 knowledge

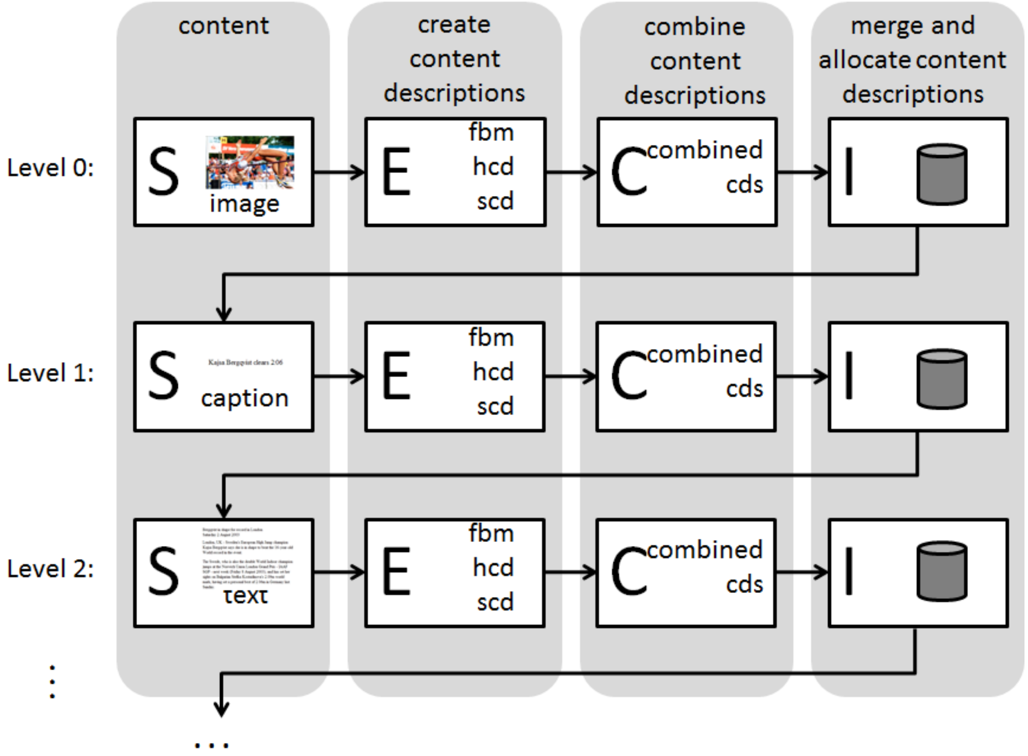
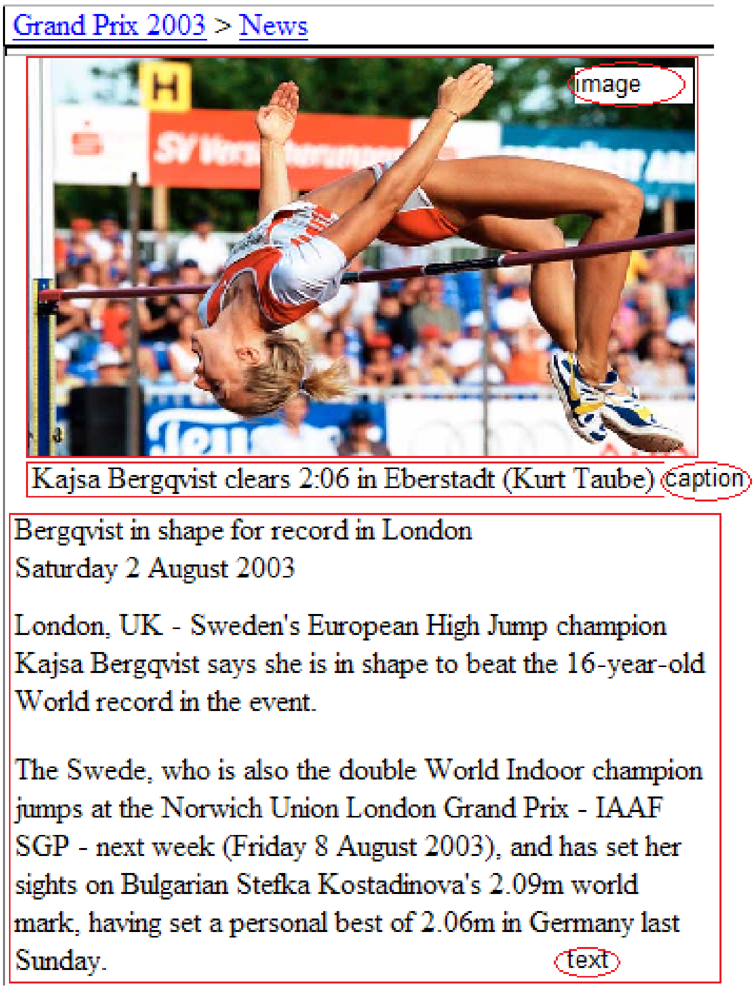


Figure 5.8: **One example of the SECI process for a multimedia document**. Description see text.

is created for a multimedia object, i.e. an *image* (pre-asset). At the socialization stage this multimedia object *image* is specified as content (ico-asset). At the externalization stage, content descriptions, for instance, feature-based metadata (fbm), holistic content descriptions (hcd), and symbolic content descriptions (scd) are created for the object *image* (log-asset). After that, all content descriptions will be combined (combination stage), and then, the content and the content descriptions (cds) is merged (com-asset) and allocated (internalization stage). The SECI process at level 1 works analogously to the first level with the difference that the input is another one: the multimedia object is *caption*. The SECI process at level 2, and so forth, works analogously to level 1.

Example 5.1 (SECI process) *The multimedia document presented in Figure 5.9 contains three multimedia objects: image, caption, and text.*



Grand Prix 2003 > News

Kajsa Bergqvist clears 2:06 in Eberstadt (Kurt Taube)

Bergqvist in shape for record in London
Saturday 2 August 2003

London, UK - Sweden's European High Jump champion Kajsa Bergqvist says she is in shape to beat the 16-year-old World record in the event.

The Swede, who is also the double World Indoor champion jumps at the Norwich Union London Grand Prix - IAAF SGP - next week (Friday 8 August 2003), and has set her sights on Bulgarian Stefka Kostadinova's 2.09m world mark, having set a personal best of 2.06m in Germany last Sunday.

Figure 5.9: **Multimedia document which contains three multimedia objects: image, caption, and text.** Adapted from [IAA09].

Pre-asset In this example, the pre-asset contains the multimedia objects (image and caption), which are presented in Figure 5.9.

Ico-asset Symbolic and holistic content representations at low-level for the pre assets of the Figure 5.9 are represented via ico assets. In our example A-boxes such as those given in Figures 4.9 and 4.10 are ico-assets.

Log-asset Symbolic and holistic content representations at high-level for the ico-assets are represented via log-assets. In our example, A-boxes such as those given in 4.15, 4.16, and 4.18 contain high-level symbolic content descriptions.

Section 4.1 gives some examples for holistic representations.

Com-asset Com-assets represents combined symbolic and holistic representations. In our example the output if the HolSym algorithm delivers com assets.

Ba *Ba* is a user-specific environment. In this thesis we use the ba environment as a context-specific repository for information retrieval tasks. If a user searches for information about jumping events all user specific queries could be predict during typing the query (i.e., via word2vec approach [MCCD13]), or a query can be stored and later used for further IR tasks. For the second case, a very simple approach could be i.e., the first string query of an user is "long jump", than the terms "long" and "jump" are stored in a context-specific repository (ba). The particular query vector for the example presented in Section 4.1 is $\vec{q} = \left\langle \begin{array}{ccccccccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right\rangle^T$.

If the second query is "high jump", the terms "long", "jump", and "high" will be considered in a so-called ba-query. Our ideas of a ba-query is that

1. the original user query of the HolSym algorithm is replaced by a ba-query or
2. the best 10 retrieval results of the original query vector $\vec{q}$ will be linked to documents which answers of query vector $\vec{q}_{\text{ba}}$.

Ba usage: Case 1 In the first case we assume that the ba-query vector $\vec{q_{ba}}$ is $\vec{q}_{\text{ba}} = \left\langle \begin{array}{ccccccccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right\rangle^T$. The query vector $\vec{q}$ is replaces of $\vec{q_{ba}}$ so that the following HolSym algorithm is solved

$$HolSym(\vec{q}_{\text{ba}}, (Docs, H, \_, \_, \_), (\_, H_{sym}, \mathcal{T}, Sym', \_), \theta, g, \delta).$$

As an example, we use the same holistic representation of documents H from Table 4.2 and the parameter $k = 2$. Then the holistic representation in a 2-dimensional space is

$$V_2'^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.04 & 0.07 & 0.99 \\ 0 & 0.01 & 0.01 & 0.02 & 0.02 & 0.01 & 0.41 & 0.27 & 0.86 & -0.13 \end{pmatrix}$$

The reduced query vector $\vec{q2_{ba}}$ for the string query is $\vec{q2_{ba}} = \langle 0.03, 0.30 \rangle$. The score S_H is presented in Table 5.15.

Docs	S_H	$S_{H_{sym}}$	g_{rel} with $g=3$	g_{rel} with $g=4$
d_1	1.00	0.92	1	2
d_2	0.93	0.93	2	1
d_3	1.00	0.98	2	3
d_4	1.00	0.97	2	3
d_5	1.00	0.88	2	2
d_6	1.00	0.96	2	3
d_7	0.95	0.36	3	4
d_8	1.00	0.13	3	4
d_9	1.00	0.97	2	3
d_{10}	0.03	0.97	2	3

Table 5.15: **Scores S_H of *Docs* and $S_{H_{sym}}$ for the ba-query "*high long jump*" using H and H_{sym} as input.**

If the threshold is $\theta = 0.95$, QA_{hol} will be:

$$docs = \{d_1, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

This result set has many false positives. Better IR results could be archived via by increasing the parameter k, decreasing the threshold, or using the HolSym Methodology.

The holistic representation in a 4-dimensional space is:

$$V_4'^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -0.15 & -0.04 & -0.07 & -0.99 \\ 0 & -0.01 & -0.01 & -0.02 & -0.02 & -0.01 & -0.41 & -0.27 & -0.86 & 0.13 \\ 0 & -0.01 & -0.01 & -0.03 & -0.02 & -0.01 & -0.12 & -0.93 & 0.35 & 0.03 \\ -0.01 & -0.02 & -0.02 & -0.05 & -0.04 & -0.02 & 0.89 & -0.25 & -0.36 & -0.10 \end{pmatrix}$$

The reduced query vector $\vec{q4_{ba}}$ for the string query is:

$$\vec{q4_{ba}} = \langle -0.03, -0.30, -0.22, -0.27 \rangle .$$

The score S is presented in Table 5.16.

Docs	S
d_1	0.95
d_2	0.91
d_3	0.96
d_4	0.95
d_5	0.94
d_6	0.94
d_7	0.13
d_8	0.76
d_9	0.71
d_{10}	0.04

Table 5.16: **Scores S of *Docs* for the ba-query "*high long jump.*"**

If the threshold is $\theta = 0.95$, QA_{hol} will be (cf. Table 5.15):

$$docs = \{d_1, d_3, d_4\}.$$

Document d_1 is a high jump document without long jump news. The documents d_3 and d_4 are documents with jumping news (i.e., hurdle run), but these false positive documents have no long jump news. If the threshold is $\theta = 0.9$, QA_{hol} will be (cf. Table 5.15):

$$docs = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9\}.$$

The result set has many false positives.

For solving QA_{sem} it is required to compute document clusters first. The operator $retrieveClusterResult(Docs, g_{rel}, S_{H_{sym}}, 3, 0.95)$ with $Docs = \langle d_1 \ldots d_{10} \rangle$, and $g_{rel} = \langle 1, 1, 1, 1, 1, 1, 3, 3, 1, 1 \rangle$ computes the following cluster:

$$Docs' = \{\{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}\}.$$

QA_{sem} returns:

$$docs = \{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}.$$

In the result set are many false positives.

The $\mathit{retrieveClusterResult}(\mathit{Docs}, g_{rel}, S_{H_{sym}}, 4, 0.95)$ with $\mathit{Docs} = \langle d_1 \dots d_{10} \rangle$, and $g_{rel} = \langle 2, 1, 3, 3, 2, 3, 4, 4, 3, 3 \rangle$, the new result set Docs' will be:

$$\mathit{Docs}' = \{\{d_3, d_4, d_6, d_9, d_{10}\}\}.$$

QA_{sem} returns:

$$\mathit{docs} = \{d_3, d_4, d_6, d_9, d_{10}\}.$$

In the result set are a little less false positives. But d_1 and d_5 are not in the result set.

If the threshold is decreased to $\theta = 0.9$, the $\mathit{retrieveClusterResult}(\mathit{Docs}$, g_{rel}, $S_{H_{sym}}$, 4, 0.95) with $\mathit{Docs} = \langle d_1 \dots d_{10} \rangle$, and the document group numbers $g_{rel} = \langle 2, 1, 3, 3, 2, 3, 4, 4, 3, 3 \rangle$ will deliver the new result set Docs':

$$\mathit{Docs}' = \{\{d_1, d_5\}\{d_3, d_4, d_6, d_9, d_{10}\}\}.$$

QA_{sem} returns (if the string query “long high jump” is transferred to the two symbolic queries $cq_1 := \{x | \mathit{HighJump}(x)\}$ and $cq_6 := \{x | \mathit{LongJump}(x)\}$):

$$\mathit{docs} = \{d_1, d_5\}.$$

Docs' represents a good clustering as well as the retrieval results because the documents d_1 and d_5 represent the same domain (athletics news). The other documents are false positive documents.

The operation $\mathit{HolSym}(\vec{q_k}, (\mathit{Docs}, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, \mathit{Sym}', \_), 0.95, 3, 3)$ returns in this example:

$$(\mathit{docs}', \mathit{docs}'') = (\{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\}, \{d_1, d_5\})$$

The operation $\mathit{SymHol}(\vec{q_k}, (\mathit{Docs}, H, \_, \_, \_), (\_, H_{sem}, \mathcal{T}, \mathit{Sym}', \_), 0.95, 3, 2)$ returns in this example:

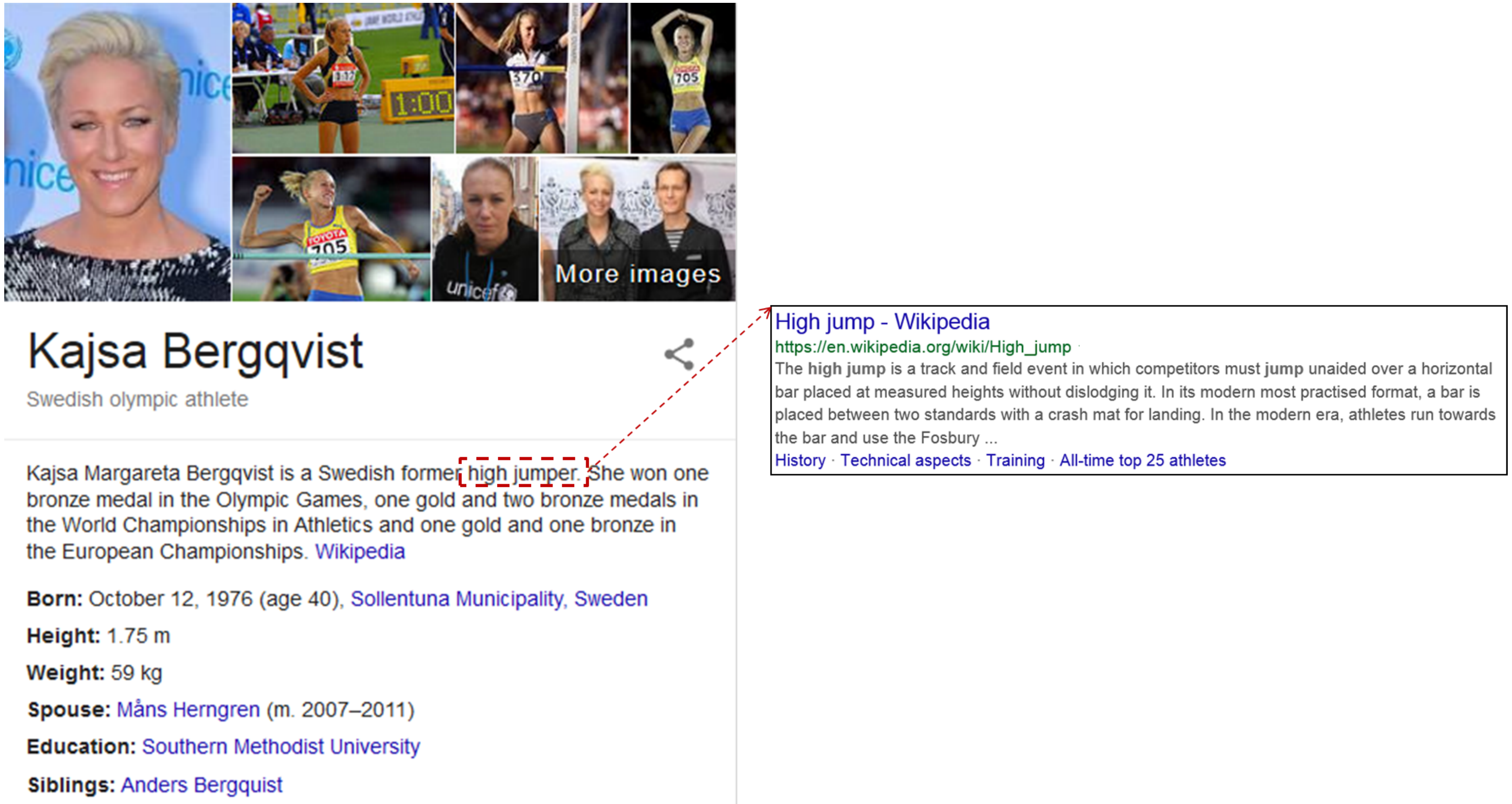
$$(\mathit{docs}', \mathit{docs}'') = (\{d_1, d_5\}, \{d_2, d_3, d_4, d_5, d_6, d_9, d_{10}\})$$

The HolSym and the SymHol algorithms deliver both high recall and precision results.

These experiments show that knowledge is created using a ba-query vector instead of the original query vector, thus we present that is possible to realize Nonaka’s concept of knowledge-creation with the extension of the ba concept.

Ba usage: Context-specific service In the second case and for our knowledge management scenario that concretely means that an extension of Google's symbolic content representation for the string query "Kajsa Bergqvist" could be created so that in our case a hyperlink to "high jumper" exists. In the following we extend our example using a ba environment.

If the first string query is "Kajsa Bergqvist", the particular query vector will be $\vec{q} = \left\langle \begin{array}{ccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right\rangle^T$. The vector is a null vector because of the missing term "Kajsa Bergqvist" in our example (cf. Table 4.2 in Section 4). In the reality (cf. Google), we can assume that there exists a term-document matrix which contains the terms "Kajsa Bergqvist" and Google's KV delivers a symbolic representation of the person Kajsa Bergqvist. But if a user types the string query "Kajsa Bergqvist high jump", Google's KV does not deliver the symbolic description presented in Figure 2.3 (right). But the text in research results have the query terms which are in bold. Our suggestion is that the terms in bold are linked with documents (see Figure 5.10). An approach how to link (highlight) documents, images, text etc. in order to support a context-specific service is presented by Espinosa in [EP11]. This kind of improvement provides the creativity mode of an engineer in a proactive way automatically.



More images

Kajsa Bergqvist

Swedish olympic athlete

Kajsa Margareta Bergqvist is a Swedish former high jumper. She won one bronze medal in the Olympic Games, one gold and two bronze medals in the World Championships in Athletics and one gold and one bronze in the European Championships. Wikipedia

Born: October 12, 1976 (age 40), Sollentuna Municipality, Sweden

Height: 1.75 m

Weight: 59 kg

Spouse: Måns Herngren (m. 2007–2011)

Education: Southern Methodist University

Siblings: Anders Bergquist

High jump - Wikipedia

https://en.wikipedia.org/wiki/High_jump

The **high jump** is a track and field event in which competitors must **jump** unaided over a horizontal bar placed at measured heights without dislodging it. In its modern most practised format, a bar is placed between two standards with a crash mat for landing. In the modern era, athletes run towards the bar and use the Fosbury ...

History · Technical aspects · Training · All-time top 25 athletes

Figure 5.10: Extended Google's symbolic content representation for the string query "Kajsa Bergqvist" w.r.t. the ba environment.

Chapter 6

Conclusion and Outlook

In this thesis we present the dependency between the quality of information retrieval results and the computation of latent structures in documents. We show that in the field of distributional semantics, e.g. LSI combined with clustering, approaches exist for finding latent structures so that the recall of retrieved documents increases. However precision decreases. Over time further holistic and symbolic retrieval algorithms or systems were improved in increasing recall and precision by computing latent structures effectively and efficiently. The improvements comprise increasing quality of holistic and symbolic document representations and performance of suggested algorithms or systems. Often such holistic and symbolic approaches were developed separately because it is hard to simultaneously maximize quality measures for query answers. In this thesis we present a new algorithm, which is called *HolSym Methodology*. The new algorithm systematically combines holistic and symbolic IR in such a way that recall and precision of retrieved documents is high.

For the holistic part of the HolSym Methodology we suggest to use LSI combined with clustering for computing latent structures of documents. Instead of the classical holistic document representation $H = V^T$, we define a new semantics representation H_{Sem} based on symbolic data as a new input parameter for LSI. After the reduction of document dimensionality with H_{Sem} as a new input parameter, the initial central location of the cluster is computed via a scoring function. Afterwards the query answering problem is solved in a way that documents with high recall and precision are determined.

The retrieval of high-quality documents is a frequent problem in KM contexts. Indeed, in [NKT98] Nonaka, Konno, and Toyama present a knowledge creation process, but this process is too informal because, in general, there only is a vague explanation how to create and share knowledge and it is not specified with what the creation is accomplished. We present that Nonaka's visionary model is practically realizable. Therefore we define holistic and symbolic operators for creating knowledge management units, which are called semantic assets, in order to have a formalized symbolic knowledge creation process.

An other important objective of this work is the systematic combination of holistic and symbolic representations. We explore the feasibility and we examined the quality of the usage of the HolSym Methodology through experimental studies, and investigate the quality in terms of recall and precision.

We identified several promising directions for future work. We have shown in [MGK$^+$14] that from a systems engineering point of view, the A-box difference operator of the HolSym Methodology has several advantages compared to alternative solutions in which existing tools such as those presented in [Fal07, Mag14] have to be enhanced in order to retrieve semantic differences of, i.e., developed components. We are planning to integrate the A-box difference operator in developing tools for automatically computing semantic differences. We believe that the A-box difference operator enhances the development of systems (cf. [MGK$^+$14]) because existing tools, i.e., in the area of model-based systems engineering do not offer reasoner's specific operations. In [BMGK15, MWHG16, AMGS17] we have identified some use cases for using the A-box difference operator.

For a contextual predictive search approach, i.e. word2vec, it is conceivable that T-boxes play a crucial role in order to predict queries in a proactive way. Nonaka describes the knowledge creating process having different levels for creating knowledge. At level 0 a holistic query answering problem increases symbolic information retrieval results during the combination process at level 1. Our idea is that we have at level 0 a string query q. The holistic query answering problem $QA_{\mathrm{hol}}(q, (Docs, H, \_, \_, \_), \theta)$ (the first part of the HolSym Algorithm) returns a set of documents $docs$. Each document has a symbolic representation D. In order to find more context-based retrieval results, at the next level a new query q' is to compute such that it holds $D' \sqsubseteq D$ (or $D' \sqsupseteq D$) and $q' \sqsubseteq q$ (or $q' \sqsupseteq q$). In addition, the new identified CGIs are added to the

T-box at the combination process for further IR tasks. We suggest to create some experiments in this field in order to have another potential to improve IR processes.

Bibliography

[ACA13] M. Nazir Ahmad, R. Colomb, and M. Abdullah. *Ontology-based Applications for Enterprise Systems & Knowledge Management.* IGI Global, 2013. [152]

[AF02] Olde B. A. and D. R. Franceschetti. The right stuff: Do you need to sanitize your corpus when using latent semantic analysis? In *IN PROCEEDINGS OF THE 24TH ANNUAL MEETING OF THE COGNITIVE SCIENCE SOCIETY. PP*, pages 708–713, 2002. [117]

[AKS06] C. Aswani Kumar and S. Srinivas. Latent semantic indexing using eigenvalue analysis for efficient information retrieval. *International Journal of Applied Mathematics and Computer Science*, 16:551–558, 2006. [117]

[AM13] H. Azgomi and A. Mahjur. A solution for calculating the false positive and false negative in lsh method to find similar documents. 3:466–472, 01 2013. [35]

[AMGS17] D. Arndt, S. Melzer, R. God, and M. Sieber. *Konzept zur Verhaltensmodellierung mit der Systems Modeling Language (SysML) zur Simulation varianten Systemverhaltens*, pages 115–124. Carl Hanser Verlag, 2017. [8, 140]

[AMVG09] J. A. Mesa Gonzalez A. Mendez-Vilas, A. Solano Martin and J. Mesa Gonzalez. *Research, Reflections and Innovations in Integrating ICT in Education.* FORMATEX, 2009. [156]

[BCM$^+$03] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, 2003. [36, 144]

[BKM] Tanya Braun, Felix Kuhr, and Ralf Möller. Unsupervised Text Annotations. In *Proceedings of the 6th Workshop on Dynamics of Knowledge and Belief (DKB-2017) and the 5th Workshop KI & Kognition (KIK-2017) co-located with 40th German Conference on Artificial Intelligence (KI 2017), series = CEUR Workshop Proceedings, volume = 1928, year = 2017, month = 25.-29.09., pages = 23–30, publisher = CEUR-WS.org, url = http://ceur-ws.org/Vol-1928/paper2.pdf.* [110]

[Ble12] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. [55]

[BMGK15] T. Bahns, S. Melzer, R. God, and D. Krause. *Ein modellbasiertes Vorgehen zur variantengerechten Entwicklung modularer Produktfamilien*, pages 141–150. Carl Hanser Verlag, 2015. [8, 140]

[BMM92] D. S. Blank, L. A. Meeden, and J. B. Marshall. Exploring the Symbolic/Subsymbolic Continuum: A Case Study of RAAM. In John Dinsmore, editor, *The Symbolic and Connectionist Paradigms: Closing the Gap*, pages 113–148. Erlbaum, Hillsdale, NJ, 1992. [2]

[BN03] F. Baader and W. Nutt. Basic description logics. In Baader et al. [BCM$^+$03], chapter 2, pages 43–95. [40]

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, pages 993–1022, March 2003. [29, 34]

[Bos08] S. Bossung. *Conceptual Content Modeling - Languages, Applications, and Systems.* PhD thesis, Hamburg University of Technology (TUHH), 2008. [6, 7, 27, 28]

[Bra08] R. B. Bradford. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of* *the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 153–162, New York, NY, USA, 2008. ACM. [113, 117]

[BRP07] R. Budiu, C. Royer, and P. Pirolli. Modeling information scent: A comparison of lsa, pmi and glsa similarity measures on common tests and corpora. In *Large scale semantic access to content (text, image, video, and sound)*, pages 314–332. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2007. [117]

[Cas23] E. Cassirer. *Philosophie der symbolischen Formen.* Number Bd. 1 in Philosophie der symbolischen Formen. B. Cassirer, 1923. [28]

[Cau13] J. Martinez Caudillo. Evaluation of fusion operators. Projektarbeit, TU Hamburg-Harburg, June 2013. [116]

[CBK$^+$10] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010. [68]

[CFO10] E. Curry, A. Freitas, and S. O'Riain. *The Role of Community-Driven Data Curation for Enterprises*, pages 25–47. Springer US, 2010. [6, 69]

[Che03] B. Cheng. Towards understanding latent semantic indexing. 2003. [117]

[CN64] E. Cassirer and H. Noack. *Philosophie der symbolischen Formen.* Number Bd. 3 in Philosophie der symbolischen Formen. Wissenschaftliche Buchgesellschaft, 1964. [27]

[CW99] J. W. Cortada and J. A. Woods, editors. *The Knowledge Management Yearbook 1999-2000.* Butterworth-Heinemann, 1999. [15]

[DACN03] M. Dierkes, A. B. Antal, J. Child, and I. Nonaka, editors. *Handbook of Organizational Learning and Knowledge.* Oxford University Press, 2003. [12, 19]

[DB05] C. De Brun. ABC of Knowledge Management. *NHS National Library for Health: Knowledge Management Specialist Library*, 2005. [15]

[DDF$^+$90a] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, (6):391–407, 1990. [3, 32, 96, 117]

[DDF$^+$90b] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. L, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990. [9, 30]

[DMG$^+$14] X. L. Dong, K. Murphy, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, T. Strohmann, and W. Zhang. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *DEXA '00: Proceedings of the 11th International Workshop on Database and Expert Systems Applications*. KDD 2014, 2014. [2, 21, 29, 35, 96]

[DSdGM15] A. Dittmar, M. Sikorski, T. de Greef, and K. Marasek, editors. *Proceedings of the European Conference on Cognitive Ergonomics 2015*. ACM, 2015. [27, 148]

[Dum03] Susan Dumais. Data-driven approaches to information access. *Cognitive Science*, 27(3):491–524, 2003. [117]

[EKM09a] S. Espinosa, A. Kaya, and R. Möller. The boemie semantic browser: A semantic application exploiting rich semantic metadata. In *Proceedings of the Applications of Semantic Technologies Workshop (AST-2009), Lübeck, Germany*, 2009. [83]

[EKM09b] S. Espinosa, A. Kaya, and R. Möller. Formalizing multimedia interpretation based on abduction over description logic aboxes. In *Proc. of the 2009 International Workshop on Description Logics DL- 2009, 27 to 30 July 2009, Oxford, United Kingdom*, 2009. CEUR Workshop Proceedings (Vol. 477). [83]

[EKS13] Thomas Eiter, Thomas Krennwallner, and Patrik Schneider. *Lightweight Spatial Conjunctive Query Answering Using Keywords*, pages 243–258. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. [108, 109]

[EP11] I.S. Espinosa PeraldÃ. *Content management and knowledge management : two faces of ontology-based deep-level interpretation of text*. PhD thesis, Hamburg University of Technology (TUHH), Hamburg, Germany, 2011. ISBN 978-3-86387-066-9. [1, 4, 136]

[ES97] J. Essers and J. Schreinemakers. Nonaka's Subjectivist Conception of Knowledge in Corporate Knowledge Program. *Knowledge Organization*, pages 24–32, 1997. [5, 20]

[Est08] J. Estefan. Survey of Candidate Model-Based Systems Engineering (MBSE) Methodologies. *International Council on Systems Engineering (INCOSE)*, pages INCOSE–TD–2007–003–02, 2008. [5]

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. [32]

[Fal07] J. Falk. Entwicklung von Differenzmetriken und deren Visualisierung. Diploma thesis, Universität Siegen, 2007. [140]

[FPBP16] Stefano Faralli, Alexander Panchenko, Chris Biemann, and Simone P Ponzetto. Linked disambiguated distributional semantic networks. In *International Semantic Web Conference*, pages 56–64. Springer, 2016. [3]

[Gee03a] K. R. Gee. Using latent semantic indexing to filter spam. In *In Proceedings of the 2003 ACM symposium on Applied computing*, pages 460–464. ACM Press, 2003. [4]

[Gee03b] Kevin R Gee. Using latent semantic indexing to filter spam. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 460–464. ACM, 2003. [117]

[Gei06] J Geiß. Latent semantic indexing and information retrieval: a quest with bosse. Master's thesis, Ruprecht-Karls University, Heidelberg, January 2006. [117]

[HB12] K. Halland and K. Britz. Abox abduction in alc using a dl tableau. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '12, pages 51–58, New York, NY, USA, 2012. ACM. [46]

[HH15] A. Hüttig and M. Herczeg. Tool-based gradual user modeling for usability engineering. In Dittmar et al. [DSdGM15], pages 11:1–11:4. [12]

[HK15] M. Herczeg and M. Koch. Allgegenwärtige Mensch-Computer-Interaktion. *Informatik-Spektrum*, 38(4):290–295, August 2015. [12, 27]

[HMW07] V. Haarslev, R. Möller, and M. Wessel. RacerPro User's Guide and Reference Manual. Version 1.9.1, February 2007. [48]

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. [34]

[Hof03] Thomas Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 259–266, New York, NY, USA, 2003. ACM. [34]

[HSD01] P. Husbands, H. Simon, and C. Ding. On the use of the singular value decomposition for text retrieval. *Computational information retrieval*, 5:145–156, 2001. [117]

[HTDRP07] D. T. Haley, P. Thomas, A. De Roeck, and M. Petre. Tuning an lsa-based assessment system for short answers in the domain of computer science: the elusive optimum dimension. *F. Wild, M. Kalz, J. van Bruggen, R. Koper (Eds.)*, page 22, 2007. [117]

[Hul94] D. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *SIGIR'94*, pages 282–291. Springer, 1994. [117]

[IAA09] IAAF (International Association of Athletics Federations), February 2009. `http://www.iaaf.org`. [55, 56, 70, 131]

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM. [35, 114]

[Ing99] J. Ingenerf. Telemedicine and terminology: Different needs of context information. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, pages 92–100, Jun 1999. [27]

[Int12] International Organization for Standardization. ISO 16684-1:2012: Graphic technology - Extensible metadata platform (XMP) specification - Part 1: Data model, serialization and core properties. `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57421`, 2012. [12]

[JH14] Sujay Kumar Jauhar and Eduard Hovy. Inducing latent semantic relations for structured distributional semantics. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 698–708, 2014. [3]

[JL00] F. Jiang and M. L. Littman. Approximate dimension equalization in vector-based information retrieval. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 423–430. Morgan Kaufmann, 2000. [117]

[JM03] E. Jessup and J. Martin. Taking a new look at the latent semantic analysis approach to information retrieval. In *Computational Info*, pages 121–144. SIAM Publishing, 2003. [117]

[Jon72] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972. [29]

[Kay11] A. Kaya. *A Logic-Based Approach to Multimedia Interpretation*. PhD thesis, Hamburg University of Technology (TUHH), Hamburg, Germany, 2011. ISBN 978-3-86664-940-8. [1, 4, 35, 48, 74, 78, 80, 83, 84, 93, 111]

[KCZ03] Yu-Seop Kim, Jeong-Ho Chang, and Byoung-Tak Zhang. An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. *Advances in Knowledge Discovery and Data Mining*, pages 566–566, 2003. [117]

[KFN09] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 61–68, New York, NY, USA, 2009. ACM. [34]

[KKH00] P. Kanerva, J. Kristoferson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of the Cognitive Science Society*, volume 1, 2000. [117]

[Kow88] R. A. Kowalski. The early years of logic programming. *Commun. ACM*, 31(1):38–43, January 1988. [42]

[KP06] A. Kontostathis and W. M. Pottenger. A framework for understanding latent semantic indexing (lsi) performance. *Information Processing & Management*, 42(1):56–73, 2006. [117]

[LD97] T.K. Landauer and S. T. Dutnais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *PSYCHOLOGICAL REVIEW*, 104(2):211–240, 1997. [117]

[LG03] J. Lin and D. Gunopulos. Dimensionality reduction by random projection and latent semantic indexing. In *In proceedings of the Text Mining Workshop, 3rd SIAM International Conference on Data Mining*, 2003. [117]

[LG14] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014. [3]

[LH11] H. Liu and O.d Hoeber. Normal distribution re-weighting for personalized web search. *Advances in Artificial Intelligence*, pages 281–284, 2011. [118]

[Lin98] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998. [3]

[LS01] M. Lizza and F. Sartoretto. A comparative analysis of lsi strategies. *Computational information retrieval*, pages 171–181, 2001. [117]

[LST07] Y. Li and J. Shawe-Taylor. Advanced learning algorithms for cross-language patent retrieval and classification. pages 1183–1199. Information Processing and Management, 2007. [117]

[Mag14] No Magic. Cameo Systems Modeler. WWW page, 2014. Available at `http://www.nomagic.com/products/cameo-systems-modeler.html`. [140]

[MB07] D. I. Martin and M. W. Berry. Mathematical foundations behind latent semantic analysis. *Handbook of latent semantic analysis*, pages 35–55, 2007. [32]

[MBW05] A. Moldovan, R. I. Boţ, and G. Wanka. Latent semantic indexing for patent documents. *International Journal of Applied Mathematics and Computer Science*, 15:551–560, 2005. [117]

[MCCD13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. [96, 99, 132]

[MCH^{+}15] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015. [68, 96]

[Mel06] S. Melzer. A content-based publish-subscribe architecture for individualized sensor data supply (in German). Master thesis, Hamburg University of Technology (TUHH), January 2006. [1, 3]

[Mel13] S. Melzer. *On the Relationship between Ontology-based and Holistic Representations in a Knowledge Management System*, chapter 17, pages 292–323. In [ACA13], 2013. [9]

[MGK^{+}14] S. Melzer, R. God, T. Kiehl, R. Möller, and M. Wessel. *Identifikation von Varianten durch Berechnung der semantischen Differenz von Modellen*, pages 279–288. Carl Hanser Verlag, 2014. [4, 8, 115, 140]

[MGXC12] Y. Ma, T. Gu, B. Xu, and L. Chang. *An ABox Abduction Algorithm for the Description Logic ALCI*, pages 125–130. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. [46]

[MLLK04] P. Mulhem, J. W. Lim, W. K. Leow, and M. S. Kankanhallo. Advances in Digital Home Photo Albums. In *Multimedia Systems and Content-based Image Retrieval*, chapter IX, pages 201–226. Idea Group Publishing, 2004. [28]

[MOH^{+}16] R. Möller, Ö. Özcep, V. Haarslev, A. Nafissi, and M. Wessel. Abductive Conjunctive Query Answering w.r.t. Ontologies. *KI - Künstliche Intelligenz*, 30(2):177–182, 2016. [48]

[Mor05] P. Moravec. Testing dimension reduction methods for text retrieval. In *DATESO*, pages 113–124, 2005. [117]

[MRS08a] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, July 2008. [2, 4, 153]

[MRS08b] C. D. Manning, P. Raghavan, and H. Schütze. *Matrix decompositions and latent semantic indexing*, pages 403–419. In [MRS08a], 1 edition, July 2008. [9, 33]

[MWHG16] S. Melzer, U. Wittke, H. Hintze, and R. God. *Physische Architekturen variantengerecht aus Funktionalen Architekturen fÃ¼r Systeme (FAS) spezifizieren*, pages 429–438. Carl Hanser Verlag, 2016. [8, 140]

[Naf13] Anahita Nafissi. *Applying Markov logics for controlling abox abduction; Die Anwendung von Markov Logik zur Steuerung der Abox-Abduktion; Die Anwendung von Markov Logik zur Steuerung der Abox-Abduktion*. PhD thesis, 2013. Advisor: Möller, Ralf; http://tubdok.tub.tuhh.de/handle/11420/1148. [48]

[NGMR17] M. Nentwig, A. Groß, M. Möller, and E. Rahm. Distributed holistic clustering on linked data. In *OTM Conferences (2)*, volume 10574 of *Lecture Notes in Computer Science*, pages 371–382. Springer, 2017. [3]

[NK98] I. Nonaka and N. Konno. The concept of 'ba': Building a foundation for knowledge creation. *California Management Review*, pages 40–54, 1998. [16, 18]

[NKT98] I. Nonaka, N. Konno, and R. Toyama. Leading Knowledge Creation: A New Framework for Dynamic Knowledge Management. In *Second Annual Knowledge Management Conference*. University of California Berkeley, 1998. [5, 7, 19, 20, 140]

[NM95] U. Nilsson and J. Maluszynski. *Logic, Programming and Prolog*. John Wiley & Sons Ltd., 2 edition, 1995. [42]

[Non08] I. Nonaka. *The Knowledge-Creating Company (Harvard Business Review Classics).* Harvard Business School Press, 2008. [5, 15, 17]

[NT95] I. Nonaka and H. Takeuchi. *The knowledge-creating company: how Japanese companies create the dynamics of innovation.* Oxford University Press, New York, 1995. [16]

[NT01] I. Nonaka and D. J. Teece, editors. *Managing Industrial Knowledge: Creation, Transfer and Utilization.* Sage Publications, Inc., Thousand Oaks, CA, USA, 2001. [18, 19]

[NT13] M. Nickel and V. Tresp. Logistic tensor factorization for multi-relational data. 06 2013. [110]

[NTB03] I. Nonaka, R. Toyama, and P. Byosière. *A Theory of Organizational Knowledge Creation: Understanding the Dynamic Process of Creating Knowledge*, pages 491–517. Oxford University Press, 2003. [16, 19]

[Pal09] G. Paliouras. Boemie final report. Technical report, Hamburg University Of Technology, 2009. Final version 1.0. [68]

[Pan70] E. Panofsky. *Meaning in the Visual Arts.* Penguin, 1970. [28]

[Pan75] E. Panofsky. *Ikonographie und Ikonologie. Eine Einführung in die Kunst der Renaissance*, pages 36–67. In [PvD75], 1975. [28]

[Pin04] B. Pincombe. Comparison of human and latent semantic analysis (lsa) judgements of pairwise document similarities for a news corpus. Technical report, DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO SCIENCES LAB, 2004. [117]

[PKM$^+$07a] S. Espinosa Peraldi, A. Kaya, S. Melzer, R. Möller, and M. Wessel. Multimedia Interpretation as Abduction. In *Proc. DL-2007: International Workshop on Description Logics*, 2007. [2]

[PKM$^+$07b] S. Espinosa Peraldi, A. Kaya, S. Melzer, R. Möller, and M. Wessel. Towards a media interpretation framework for the semantic web. *The 2007 IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, 2007. [2, 9]

[PKMM08] S. Epinosa Peraldi, A. Kaya, S. Melzer, and R. Möller. On ontology based abduction for text interpretation. In A. Gelbuhk, editor, *Proc. of 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, number 4919 in LNCS, pages 194–205. International Society of the Learning Sciences, 2008. [8, 46, 48]

[PL07a] S. Pado and M. Lapata. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199, 2007. [3]

[PL07b] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007. [3]

[PRF$^+$17] Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. Unsupervised does not mean uninterpretable: the case for word sense induction and disambiguation. Association for Computational Linguistics, 2017. [3]

[PvD75] E. Panofsky and W. Höck von Dumont. *Sinn und Deutung in der bildenden Kunst.* Dumont, 1975. [154]

[Rac08] Shaina Race. Data Clustering via Dimension Reduction and Algorithm Aggregation. Master thesis, Graduate Faculty of North Carolina State University, 2008. [99]

[RB12] Martin Riedl and Chris Biemann. Text segmentation with topic models. *Journal for Language Technology and Computational Linguistics*, 27(1):47–69, 2012. [3]

[RG65] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. [3]

[RM10] Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010. [3]

[Sal83] Salton, G. and Fox, E. A. and Harry, W. Extended Boolean information retrieval. `https://en.wikipedia.org/wiki/Extended_Boolean_model`, October 1983. [3, 99]

[Sch03] P. Schütt. The post-Nonaka Knowledge Management. *Journal of Universal Computer Science*, pages 451–462, 2003. [15, 16]

[Sch09] J. W. Schmidt. Conceptual modeling: Foundations and applications. chapter On Conceptual Content Management, pages 153–172. Springer-Verlag, Berlin, Heidelberg, 2009. [28]

[SCRP09] N. S. A. Silva, G.J.M. Costa, S. Rogerson, and M. Prior. *Knowledge or content? The philosophical boundaries in e-learning pedagogical theories*, pages 221–225. In [AMVG09], 2009. [13]

[Seh04] H.-W. Sehring. *Konzeptorientiertes Content Management: Modell, Systemarchitektur und Prototypen.* PhD thesis, Hamburg University of Technology (TUHH), 2004. [6, 7, 11, 13]

[Sha05] M. Shanahan. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29(1):103–134, 2005. [46]

[SI09] L. Skorkovska and P. Ircing. Experiments with Automatic Query Formulation in the Extended Boolean Model. *Springer Berlin/ Heidelberg*, 2009. [3]

[SKI16] Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki. On approximately searching for similar word embeddings. In *ACL (1)*, 2016. [3]

[SLD96] I. Syu, S.-D. Lang, and N. Deo. Incorporating latent semantic indexing into a neural network model for information retrieval. In

Proceedings of the fifth international conference on Information and knowledge management, pages 145–153. ACM, 1996. [117]

[SM86] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA, 1986. [3, 29, 99]

[SMD12] B. Sarayreh, A. Mardawi, and R. Dmour. Comparative Study: The Nonaka Model of Knowledge Management. *International Journal of Engineering and Advanced Technology (IJEAT)*, pages 45–48, 2012. [5, 14, 16, 17, 18, 20]

[Sob13] E. Sober. *Core Questions in Philosophy: A Text with Readings.* Pearson Education, 2013. [46]

[SS00] U. Schneider and G. Schreyögg. *Management als Steuerung des organisatorischen Wissens.*, pages 79–110. Duncker & Humblot, Berlin, 2000. [15]

[SS03] J. W. Schmidt and H.-W. Sehring. Conceptual Content Modeling and Management: The Rationale of an Asset Language. In *Perspectives of System Informatics*, volume 2890 of *LNCS*, pages 469–493. Springer, 2003. [6, 13, 27, 28]

[SS04] H.-W. Sehring and J. W. Schmidt. Beyond Databases: An Asset Language for Conceptual Content Management. In *Proceedings of the 8th East European Conference on Advances in Databases and Information Systems*, volume 3255 of *LNCS*, pages 99–112. Springer-Verlag, 2004. [6, 12, 13, 14]

[Sun06] Ron Sun. *The CLARION cognitive architecture: Extending cognitive modeling to social simulation*, pages 79–99. Cambridge University Press, 2006. [6]

[SWY75] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, pages 613–620, November 1975. [2, 29]

[TL03] P. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003. [117]

[Ull85] J. D. Ullman. Implementation of logical query languages for databases. *ACM Trans. Database Syst.*, pages 289–321, September 1985. [43]

[VEK76] M. H. Van Emden and R. A. Kowalski. The semantics of predicate logic as a programming language. *J. ACM*, pages 733–742, October 1976. [43]

[VH05] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, September 2005. [2]

[W3C14a] W3C Recommendation. HTML+RDFa 1.1. www.w3.org/TR/rdfa-in-html/, October 2014. [6, 29]

[W3C14b] W3C Recommendation. The RDF Data Cube Vocabulary. `https://www.w3.org/TR/vocab-data-cube/`, January 2014. [12]

[Wan11] S. Wandelt. *Efficient Instance Retrieval over Semi-Expressive Ontologies*. PhD thesis, Technische Universit"at Hamburg-Harburg, 2011. [113]

[WC16] W. Y. Wang and W. W. Cohen. Learning first-order logic embeddings via matrix factorization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2132–2138. AAAI Press, 2016. [3, 99]

[WH00] P. Wiemer-Hastings. Adding syntactic information to lsa. In *Proceedings of the Cognitive Science Society*, volume 1, 2000. [117]

[WHG99] K. Wiemer-Hastings and A. C. Graesser. Improving an intelligent tutor's comprehension of students with latent semantic analysis. In *Artificial Intelligence in Education*, pages 535–542. IOS Press, 1999. [117]

[Wii86] K. Wiig. Knowledge Based Systems. In *Expert Systems: Impacts & Potentials - Proceedings of the International conference held in London*, 1986. [15]

[Wik15a] Wikipedia. Olympiastadion München. WWW page, October 2015. Available at `https://de.wikipedia.org/wiki/Olympiastadion_M%C3%BCnchen`. [21]

[Wik15b] Wikipedia. Prince Charming. WWW page, April 2015. Available at `http://en.wikipedia.org/wiki/Prince_Charming`. [56]

[Wik15c] Wikipedia. Rapunzel. WWW page, April 2015. Available at `http://en.wikipedia.org/wiki/Rapunzel`. [56]

[Wik15d] Wikipedia. sleeping Beauty. WWW page, April 2015. Available at `http://en.wikipedia.org/wiki/Sleeping_Beauty`. [56]

[Wik15e] Wikipedia. Snow White. WWW page, April 2015. Available at `http://en.wikipedia.org/wiki/Snow_White`. [56]

[WMC13] W. Y. Wang, K. Mazaitis, and W. W. Cohen. Programming with personalized pagerank: A locally groundable first-order probabilistic logic. *CoRR*, abs/1305.2254, 2013. [99]

[WMC14] W. Y. Wang, K. Mazaitis, and W. W. Cohen. Structure learning via parameter learning. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1199–1208, New York, NY, USA, 2014. ACM. [99]

[YCBF98] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: learning from bilingual corpora. *Artificial intelligence*, 103(1-2):323–345, 1998. [117]

[YYT05] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–265. ACM, 2005. [117]

[Zha15] C. Zhang. *DeepDive: a data management system for automatic knowledge base construction.* PhD thesis, The University of Wisconsin-Madison, 2015. [96]

[ZMS98] H. Zha, O. Marques, and H. Simon. *A subspace-based model for information retrieval with applications in latent semantic indexing.* Pennsylvania State University, Department of Computer Science and Engineering, College of Engineering, 1998. [117]

Appendix A

A-boxes

Annotation of Document d_7

PC-domain	:	*FairyTale*,			
$PC\text{-}IND_{1}$	:	*City*,	($PC\text{-}IND_{1}$, “Fabletown”)	:	*hasCityName*
$PC\text{-}IND_{2}$	:	*City*,	($PC\text{-}IND_{2}$, “New York City”)	:	*hasCityName*
$PC\text{-}IND_{3}$	:	*Company*,	($PC\text{-}IND_{3}$, “Rumpelstiltskin”)	:	*hasCompanyName*
$PC\text{-}IND_{4}$	:	*Entertainment*,	($PC\text{-}IND_{4}$, “Oscar”)	:	*hasEntertainment*
$PC\text{-}IND_{5}$	:	*Movie*,	($PC\text{-}IND_{5}$, “Cinderella”)	:	*hasMovieName*
$PC\text{-}IND_{6}$	:	*Movie*,	($PC\text{-}IND_{6}$, “Into the Woods”)	:	*hasMovieName*
$PC\text{-}IND_{7}$	:	*Movie*,	($PC\text{-}IND_{7}$, “Prince Charming”)	:	*hasMovieName*
$PC\text{-}IND_{8}$	:	*Movie*,	($PC\text{-}IND_{8}$, “Shrek 2”)	:	*hasMovieName*
$PC\text{-}IND_{9}$	:	*Movie*,	($PC\text{-}IND_{9}$, “Shrek the Third”)	:	*hasMovieName*
$PC\text{-}IND_{10}$	:	*Movie*,	($PC\text{-}IND_{10}$, “Sleeping Beauty”)	:	*hasMovieName*
$PC\text{-}IND_{11}$	:	*Movie*,	($PC\text{-}IND_{11}$, “Snow White”)	:	*hasMovieName*
$PC\text{-}IND_{12}$	:	*Movie*,	($PC\text{-}IND_{12}$, “Snow White and the Seven Dwarfs”)	:	*hasMovieName*
$PC\text{-}IND_{13}$	:	*Movie*,	($PC\text{-}IND_{13}$, “The Blue Bird”)	:	*hasMovieName*
$PC\text{-}IND_{14}$	:	*MusicAlbum*,	($PC\text{-}IND_{14}$, “Beautiful”)	:	*hasMusicAlbumName*
$PC\text{-}IND_{15}$	:	*MusicAlbum*,	($PC\text{-}IND_{15}$, “Fine”)	:	*hasMusicAlbumName*
$PC\text{-}IND_{16}$	:	*MusicAlbum*,	($PC\text{-}IND_{16}$, “Prince Charming”)	:	*hasMusicAlbumName*
$PC\text{-}IND_{17}$	:	*Person*,	($PC\text{-}IND_{17}$, “Abigail”)	:	*hasPersonName*
$PC\text{-}IND_{18}$	:	*Person*,	($PC\text{-}IND_{18}$, “Andrew Lang”)	:	*hasPersonName*
$PC\text{-}IND_{19}$	:	*Person*,	($PC\text{-}IND_{19}$, “Andy Lau”)	:	*hasPersonName*
$PC\text{-}IND_{20}$	:	*Person*,	($PC\text{-}IND_{20}$, “Charles Perrault”)	:	*hasPersonName*
$PC\text{-}IND_{21}$	:	*Person*,	($PC\text{-}IND_{21}$, “Chris Colfer”)	:	*hasPersonName*
$PC\text{-}IND_{22}$	:	*Person*,	($PC\text{-}IND_{22}$, “Clara”)	:	*hasPersonName*
$PC\text{-}IND_{23}$	:	*Person*,	($PC\text{-}IND_{23}$, “David Charvet”)	:	*hasPersonName*
$PC\text{-}IND_{24}$	:	*Person*,	($PC\text{-}IND_{24}$, “David Nolan”)	:	*hasPersonName*
$PC\text{-}IND_{25}$	:	*Person*,	($PC\text{-}IND_{25}$, “Giselle”)	:	*hasPersonName*
$PC\text{-}IND_{26}$	:	*Person*,	($PC\text{-}IND_{26}$, “James”)	:	*hasPersonName*
$PC\text{-}IND_{27}$	:	*Person*,	($PC\text{-}IND_{27}$, “Josh Dallas”)	:	*hasPersonName*
$PC\text{-}IND_{28}$	:	*Person*,	($PC\text{-}IND_{28}$, “Ling Ling”)	:	*hasPersonName*
$PC\text{-}IND_{29}$	:	*Person*,	($PC\text{-}IND_{29}$, “Mary Margaret”)	:	*hasPersonName*
$PC\text{-}IND_{30}$	:	*Person*,	($PC\text{-}IND_{30}$, “Prince”)	:	*hasPersonName*
$PC\text{-}IND_{31}$	:	*Person*,	($PC\text{-}IND_{31}$, “Robert Scott”)	:	*hasPersonName*
$PC\text{-}IND_{32}$	:	*Person*,	($PC\text{-}IND_{32}$, “Robert Sheckley”)	:	*hasPersonName*
$PC\text{-}IND_{33}$	:	*Person*,	($PC\text{-}IND_{33}$, “Roger Zelazny”)	:	*hasPersonName*
$PC\text{-}IND_{34}$	:	*Person*,	($PC\text{-}IND_{34}$, “Tia Carrere”)	:	*hasPersonName*
$PC\text{-}IND_{35}$	:	*Position*,	($PC\text{-}IND_{35}$, “Head of Prince”)	:	*hasPositionName*
$PC\text{-}IND_{36}$	:	*Position*,	($PC\text{-}IND_{36}$, “King”)	:	*hasPositionName*
$PC\text{-}IND_{37}$	:	*Position*,	($PC\text{-}IND_{37}$, “Prince”)	:	*hasPositionName*
$PC\text{-}IND_{38}$	:	*Position*,	($PC\text{-}IND_{38}$, “mayor”)	:	*hasPositionName*
$PC\text{-}IND_{39}$	:	*Position*,	($PC\text{-}IND_{39}$, “princess”)	:	*hasPositionName*
$PC\text{-}IND_{40}$	:	*Position*,	($PC\text{-}IND_{40}$, “young actress”)	:	*hasPositionName*
$PC\text{-}IND_{41}$	:	*TVShow*,	($PC\text{-}IND_{41}$, “Once Upon a Time”)	:	*hasTVShowName*

$PC\text{-}IND_{42}$	:	*TVShow*,	($PC\text{-}IND_{42}$, “Prince Charming”)	:	*hasTVShowName*
$PC\text{-}IND_{43}$	:	*Technology*,	($PC\text{-}IND_{43}$, “Adam”)	:	*hasTechnologyName*

Table A.1: Analysis Abox $\mathcal{A}_{d_7}$

Annotation of Document d_8

(*R-domain*	:	*FairyTale*),			
($R\text{-}IND_{1}$	:	*City*),	(($R\text{-}IND_{1}$, Dresden)	:	*hasCityName*)
($R\text{-}IND_{2}$	:	*City*),	(($R\text{-}IND_{2}$, Helsinki)	:	*hasCityName* “”))
($R\text{-}IND_{3}$	:	*City*),	(($R\text{-}IND_{3}$, Leipzig)	:	*hasCityName*)
($R\text{-}IND_{4}$	:	*City*),	(($R\text{-}IND_{4}$, New Orleans)	:	*hasCityName*)
($R\text{-}IND_{5}$	:	*Company*),	(($R\text{-}IND_{5}$, Delphi)	:	*hasCompanyName*)
($R\text{-}IND_{6}$	:	*Company*),	(($R\text{-}IND_{6}$, Oracle)	:	*hasCompanyName*)
($R\text{-}IND_{7}$	:	*Company*),	(($R\text{-}IND_{7}$, Princeton University Press)	:	*hasCompanyName*)
($R\text{-}IND_{8}$	:	*Company*),	(($R\text{-}IND_{8}$, The Walt Disney Company)	:	*hasCompanyName*)
($R\text{-}IND_{9}$	:	*Company*),	(($R\text{-}IND_{9}$, W W Norton & Company Incorporated)	:	*hasCompanyName*)
($R\text{-}IND_{10}$	:	*Country*),	(($R\text{-}IND_{10}$, France)	:	*hasCompanyName*)
($R\text{-}IND_{11}$	:	*Country*),	(($R\text{-}IND_{11}$, Germany)	:	*hasCompanyName*)
($R\text{-}IND_{12}$	:	*Country*),	(($R\text{-}IND_{12}$, United States)	:	*hasCompanyName*)
($R\text{-}IND_{13}$	:	*Facility*),	(($R\text{-}IND_{13}$,The Tower)	:	*hasFacilityName*)
($R\text{-}IND_{14}$	:	*IndustryTerm*),	(($R\text{-}IND_{14}$, food)	:	*hasIndustryTermName*)
($R\text{-}IND_{15}$	:	*IndustryTerm*),	(($R\text{-}IND_{15}$, media)	:	*hasIndustryTermName*)
($R\text{-}IND_{16}$	:	*Movie*),	(($R\text{-}IND_{16}$, Beauty and the Beast)	:	*hasMovieName*)
($R\text{-}IND_{17}$	:	*Movie*),	(($R\text{-}IND_{17}$, Grimm)	:	*hasMovieName*)
($R\text{-}IND_{18}$	:	*Movie*),	(($R\text{-}IND_{18}$, Into the Woods)	:	*hasMovieName*)
($R\text{-}IND_{19}$	:	*Movie*),	(($R\text{-}IND_{19}$, Persinette)	:	*hasMovieName*)
($R\text{-}IND_{20}$	:	*Movie*),	(($R\text{-}IND_{20}$, Shrek the Third)	:	*hasMovieName*)
($R\text{-}IND_{21}$	:	*Movie*),	(($R\text{-}IND_{21}$, Tangled)	:	*hasMovieName*)
($R\text{-}IND_{22}$	:	*Movie*),	(($R\text{-}IND_{22}$, The Blue Bird)	:	*hasMovieName*)
($R\text{-}IND_{23}$	:	*MusicGroup*),	(($R\text{-}IND_{23}$, Brothers Grimm)	:	*hasMusicGroupName*)
($R\text{-}IND_{24}$	:	*Organization*),	(($R\text{-}IND_{24}$, Princeton University)	:	*hasOrganizationName*)
($R\text{-}IND_{25}$	:	*Person*),	(($R\text{-}IND_{25}$, Andrew Lang)	:	*hasPersonName*)
($R\text{-}IND_{26}$	:	*Person*),	(($R\text{-}IND_{26}$, Annotated Rapunzel)	:	*hasPersonName*)
($R\text{-}IND_{27}$	:	*Person*),	(($R\text{-}IND_{27}$, Charlotte)	:	*hasPersonName*)
($R\text{-}IND_{28}$	:	*Person*),	(($R\text{-}IND_{28}$,Donna Murphy)	:	*hasPersonName*)
($R\text{-}IND_{29}$	:	*Person*),	(($R\text{-}IND_{29}$, Eurydice)	:	*hasPersonName*)
($R\text{-}IND_{30}$	:	*Person*),	(($R\text{-}IND_{30}$, Fiona)	:	*hasPersonName*)
($R\text{-}IND_{31}$	:	*Person*),	(($R\text{-}IND_{31}$, Friedrich Schulz Kleine Romane)	:	*hasPersonName*)
($R\text{-}IND_{32}$	:	*Person*),	(($R\text{-}IND_{32}$, Gena Rowlands)	:	*hasPersonName*)
($R\text{-}IND_{33}$	:	*Person*),	(($R\text{-}IND_{33}$, Gothel)	:	*hasPersonName*)
($R\text{-}IND_{34}$	:	*Person*),	(($R\text{-}IND_{34}$, Heidi Anne)	:	*hasPersonName*)
($R\text{-}IND_{35}$	:	*Person*),	(($R\text{-}IND_{35}$, Jack Zipes)	:	*hasPersonName*)
($R\text{-}IND_{36}$	:	*Person*),	(($R\text{-}IND_{36}$, Jeff Bridges)	:	*hasPersonName*)
($R\text{-}IND_{37}$	:	*Person*),	(($R\text{-}IND_{37}$, Johnny Gruelle)	:	*hasPersonName*)
($R\text{-}IND_{38}$	:	*Person*),	(($R\text{-}IND_{38}$, Mandy Moore)	:	*hasPersonName*)
($R\text{-}IND_{39}$	:	*Person*),	(($R\text{-}IND_{39}$, Maria Tatar)	:	*hasPersonName*)
($R\text{-}IND_{40}$	:	*Person*),	(($R\text{-}IND_{40}$, Maya Rudolph)	:	*hasPersonName*)
($R\text{-}IND_{41}$	:	*Person*),	(($R\text{-}IND_{41}$, Olivia Newton)	:	*hasPersonName*)
($R\text{-}IND_{42}$	:	*Person*),	(($R\text{-}IND_{42}$, Paul O. Zelinsky)	:	*hasPersonName*)
($R\text{-}IND_{43}$	:	*Person*),	(($R\text{-}IND_{43}$, Rose de Caumont La)	:	*hasPersonName*)
($R\text{-}IND_{44}$	:	*Person*),	(($R\text{-}IND_{44}$, Ruth Manning)	:	*hasPersonName*)
($R\text{-}IND_{45}$	:	*Person*),	(($R\text{-}IND_{45}$, Shelley Duvall)	:	*hasPersonName*)
($R\text{-}IND_{46}$	:	*Person*),	(($R\text{-}IND_{46}$, Whoopi Goldberg)	:	*hasPersonName*)
($R\text{-}IND_{47}$	:	*Person*),	(($R\text{-}IND_{47}$, Zachary Levi)	:	*hasPersonName*)
($R\text{-}IND_{48}$	:	*Position*),	(($R\text{-}IND_{48}$, King)	:	*hasPositionName*)
($R\text{-}IND_{49}$	:	*Position*),	(($R\text{-}IND_{49}$, Maid)	:	*hasPositionName*)
($R\text{-}IND_{50}$	:	*Position*),	(($R\text{-}IND_{50}$, Queen)	:	*hasPositionName*)
($R\text{-}IND_{51}$	:	*Position*),	(($R\text{-}IND_{51}$, artist)	:	*hasPositionName*)
($R\text{-}IND_{52}$	:	*Position*),	(($R\text{-}IND_{52}$, author)	:	*hasPositionName*)
($R\text{-}IND_{53}$	:	*Position*),	(($R\text{-}IND_{53}$, illustrator)	:	*hasPositionName*)
($R\text{-}IND_{54}$	:	*Position*),	(($R\text{-}IND_{54}$, kitchen maid)	:	*hasPositionName*)
($R\text{-}IND_{55}$	:	*Position*),	(($R\text{-}IND_{55}$, prince)	:	*hasPositionName*)
($R\text{-}IND_{56}$	:	*Position*),	(($R\text{-}IND_{56}$, princess)	:	*hasPositionName*)
($R\text{-}IND_{57}$	:	*ProgrLanguage*),	(($R\text{-}IND_{57}$, php)	:	*hasProgrLanguageName*)
($R\text{-}IND_{58}$	:	*TVShow*),	(($R\text{-}IND_{58}$, Faerie Tale Theatre)	:	*hasTVShowName*)

($R\text{-}IND_{59}$	:	*TVShow*),	(($R\text{-}IND_{59}$, Grimm's Fairy Tale Classics)	:	*hasTVShowName*)
($R\text{-}IND_{60}$	:	*Technology*),	(($R\text{-}IND_{60}$, Animation)	:	*hasTechnologyName*)
($R\text{-}IND_{61}$	:	*TVShow*),	(($R\text{-}IND_{61}$, Happily Ever After: Fairy Tales for Every Child)	:	*hasTVShowName*)

Table A.2: Analysis Abox $\mathcal{A}_{d_8}$

Annotation of Document d_9

($SB\text{-}domain$	:	$FairyTale$),			
($SB\text{-}IND_{1}$	:	$City$),	(($SB\text{-}IND_{1}$, Paris)	:	hasCityName)
($SB\text{-}IND_{2}$	:	$Company$),	(($SB\text{-}IND_{2}$, ABC)	:	hasCompanyName)
($SB\text{-}IND_{3}$	:	$Company$),	(($SB\text{-}IND_{3}$, Mattel)	:	hasCompanyName)
($SB\text{-}IND_{4}$	:	$Company$),	(($SB\text{-}IND_{4}$, The Glass Coffin Rip Van Winkle)	:	hasCompanyName)
($SB\text{-}IND_{5}$	:	$Company$),	(($SB\text{-}IND_{5}$, Walt Disney)	:	hasCompanyName)
($SB\text{-}IND_{6}$	:	$Company$),	(($SB\text{-}IND_{6}$, the Walt Disney Company)	:	hasCompanyName)
($SB\text{-}IND_{7}$	:	$Country$),	(($SB\text{-}IND_{7}$, Germany)	:	hasCountryName)
($SB\text{-}IND_{8}$	:	$Country$),	(($SB\text{-}IND_{8}$, United Kingdom)	:	hasCountryName)
($SB\text{-}IND_{9}$	:	$Facility$),	(($SB\text{-}IND_{9}$, Jenny Harbour)	:	hasFacilityName)
($SB\text{-}IND_{10}$	:	$Facility$),	(($SB\text{-}IND_{10}$, Rosamund's castle)	:	hasFacilityName)
($SB\text{-}IND_{11}$	:	$IndustryTerm$),	(($SB\text{-}IND_{11}$, food)	:	hasIndustryTermName)
($SB\text{-}IND_{12}$	:	$IndustryTerm$),	(($SB\text{-}IND_{12}$, literature portal)	:	hasIndustryTermName)
($SB\text{-}IND_{13}$	:	$IndustryTerm$),	(($SB\text{-}IND_{13}$, mother-in-law attempting)	:	hasIndustryTermName)
($SB\text{-}IND_{14}$	:	$Movie$),	(($SB\text{-}IND_{14}$, Prinsessa Ruusunen)	:	hasMovieName)
($SB\text{-}IND_{15}$	:	$Movie$),	(($SB\text{-}IND_{15}$, Rosebud)	:	hasMovieName)
($SB\text{-}IND_{16}$	:	$Movie$),	(($SB\text{-}IND_{16}$, Sleeping Beauty)	:	hasMovieName)
($SB\text{-}IND_{17}$	:	$Movie$),	(($SB\text{-}IND_{17}$, The Brothers Grimm)	:	hasMovieName)
($SB\text{-}IND_{18}$	:	$Movie$),	(($SB\text{-}IND_{18}$, The League of Extraordinary Gentlemen)	:	hasMovieName)
($SB\text{-}IND_{19}$	:	$Movie$),	(($SB\text{-}IND_{19}$, The Rose and the Ring)	:	hasMovieName)
($SB\text{-}IND_{20}$	:	$Movie$),	(($SB\text{-}IND_{20}$, The Sleeping Beauty)	:	hasMovieName)
($SB\text{-}IND_{21}$	:	$MusicAlbum$),	(($SB\text{-}IND_{21}$, A Kiss In Time)	:	hasMusicAlbumName)
($SB\text{-}IND_{22}$	:	$MusicAlbum$),	(($SB\text{-}IND_{22}$, Moon)	:	hasMusicAlbumName)
($SB\text{-}IND_{23}$	:	$MusicAlbum$),	(($SB\text{-}IND_{23}$, Sleeping Beauty Wakes)	:	hasMusicAlbumName)
($SB\text{-}IND_{24}$	:	$MusicAlbum$),	(($SB\text{-}IND_{24}$, Sun)	:	hasMusicAlbumName)
($SB\text{-}IND_{25}$	:	$NaturalFeature$),	(($SB\text{-}IND_{25}$, Ardennes forests)	:	hasNaturalFeatureName)
($SB\text{-}IND_{26}$	:	$NaturalFeature$),	(($SB\text{-}IND_{26}$, Zellandine falls)	:	hasNaturalFeatureName)
($SB\text{-}IND_{27}$	:	$Organization$),	(($SB\text{-}IND_{27}$, League of Extraordinary Gentlemen)	:	hasOrganizationName)
($SB\text{-}IND_{28}$	:	$Organization$),	(($SB\text{-}IND_{28}$, UN Court)	:	hasOrganizationName)
($SB\text{-}IND_{29}$	:	$Person$),	(($SB\text{-}IND_{29}$, Abby Dobson)	:	hasPersonName)
($SB\text{-}IND_{30}$	:	$Person$),	(($SB\text{-}IND_{30}$, Alex Flinn)	:	hasPersonName)
($SB\text{-}IND_{31}$	:	$Person$),	(($SB\text{-}IND_{31}$, Alexander Zick)	:	hasPersonName)
($SB\text{-}IND_{32}$	:	$Person$),	(($SB\text{-}IND_{32}$, Angelina Jolie)	:	hasPersonName)
($SB\text{-}IND_{33}$	:	$Person$),	(($SB\text{-}IND_{33}$, Anne Rice)	:	hasPersonName)
($SB\text{-}IND_{34}$	:	$Person$),	(($SB\text{-}IND_{34}$, Archie Campbell)	:	hasPersonName)
($SB\text{-}IND_{35}$	:	$Person$),	(($SB\text{-}IND_{35}$, Arthur Rackham)	:	hasPersonName)
($SB\text{-}IND_{36}$	:	$Person$),	(($SB\text{-}IND_{36}$, Bois Dormant)	:	hasPersonName)
($SB\text{-}IND_{37}$	:	$Person$),	(($SB\text{-}IND_{37}$, Briar Rose)	:	hasPersonName)
($SB\text{-}IND_{38}$	:	$Person$),	(($SB\text{-}IND_{38}$, Charles Perrault)	:	hasPersonName)
($SB\text{-}IND_{39}$	:	$Person$),	(($SB\text{-}IND_{39}$, David Irving)	:	hasPersonName)
($SB\text{-}IND_{40}$	:	$Person$),	(($SB\text{-}IND_{40}$, Edward Burne-Jones)	:	hasPersonName)
($SB\text{-}IND_{41}$	:	$Person$),	(($SB\text{-}IND_{41}$, Edward Frederick Brewtnall)	:	hasPersonName)
($SB\text{-}IND_{42}$	:	$Person$),	(($SB\text{-}IND_{42}$, Emily Smith Michele Carafa)	:	hasPersonName)
($SB\text{-}IND_{43}$	:	$Person$),	(($SB\text{-}IND_{43}$, Eugene Scribe)	:	hasPersonName)
($SB\text{-}IND_{44}$	:	$Person$),	(($SB\text{-}IND_{44}$, Florimund)	:	hasPersonName)
($SB\text{-}IND_{45}$	:	$Person$),	(($SB\text{-}IND_{45}$, Fritz Genschow)	:	hasPersonName)
($SB\text{-}IND_{46}$	:	$Person$),	(($SB\text{-}IND_{46}$, Gustave DorÃ©)	:	hasPersonName))
($SB\text{-}IND_{47}$	:	$Person$),	(($SB\text{-}IND_{47}$, Hee Haw)	:	hasPersonName)
($SB\text{-}IND_{48}$	:	$Person$),	(($SB\text{-}IND_{48}$, Ivan Vsevolozhsky)	:	hasPersonName)
($SB\text{-}IND_{49}$	:	$Person$),	(($SB\text{-}IND_{49}$, Ivy Green)	:	hasPersonName)
($SB\text{-}IND_{50}$	:	$Person$),	(($SB\text{-}IND_{50}$, Jane Yolen)	:	hasPersonName)
($SB\text{-}IND_{51}$	:	$Person$),	(($SB\text{-}IND_{51}$, Jim C. Hines)	:	hasPersonName)
($SB\text{-}IND_{52}$	:	$Person$),	(($SB\text{-}IND_{52}$, Johann Georg van Caspel)	:	hasPersonName)
($SB\text{-}IND_{53}$	:	$Person$),	(($SB\text{-}IND_{53}$, John Stejean)	:	hasPersonName)
($SB\text{-}IND_{54}$	:	$Person$),	(($SB\text{-}IND_{54}$, Joseph Jacobs)	:	hasPersonName)
($SB\text{-}IND_{55}$	:	$Person$),	(($SB\text{-}IND_{55}$, Julian Morris)	:	hasPersonName)
($SB\text{-}IND_{56}$	:	$Person$),	(($SB\text{-}IND_{56}$, Kristin Bauer van Straten)	:	hasPersonName)
($SB\text{-}IND_{57}$	:	$Person$),	(($SB\text{-}IND_{57}$, Louis Sussmann-Hellborn)	:	hasPersonName)

($SB\text{-}IND_{58}$	:	*Person*),	(($SB\text{-}IND_{58}$, Mercedes Lackey)	:	hasPersonName)
($SB\text{-}IND_{59}$	:	*Person*),	(($SB\text{-}IND_{59}$, Mother)	:	hasPersonName)
($SB\text{-}IND_{60}$	:	*Person*),	(($SB\text{-}IND_{60}$, Orson Scott Card)	:	hasPersonName)
($SB\text{-}IND_{61}$	:	*Person*),	(($SB\text{-}IND_{61}$, Phillip)	:	hasPersonName)
($SB\text{-}IND_{62}$	:	*Person*),	(($SB\text{-}IND_{62}$, Prince)	:	hasPersonName)
($SB\text{-}IND_{63}$	:	*Person*),	(($SB\text{-}IND_{63}$, Rachel Sheinkin)	:	hasPersonName)
($SB\text{-}IND_{64}$	:	*Person*),	(($SB\text{-}IND_{64}$, Richard Wagner)	:	hasPersonName)
($SB\text{-}IND_{65}$	:	*Person*),	(($SB\text{-}IND_{65}$, Robert Schumann)	:	hasPersonName)
($SB\text{-}IND_{66}$	:	*Person*),	(($SB\text{-}IND_{66}$, Robin McKinley)	:	hasPersonName)
($SB\text{-}IND_{67}$	:	*Person*),	(($SB\text{-}IND_{67}$, Sailor Moon)	:	hasPersonName)
($SB\text{-}IND_{68}$	:	*Person*),	(($SB\text{-}IND_{68}$, Sarah Bolger)	:	hasPersonName)
($SB\text{-}IND_{69}$	:	*Person*),	(($SB\text{-}IND_{69}$, Silver Millennium)	:	hasPersonName)
($SB\text{-}IND_{70}$	:	*Person*),	(($SB\text{-}IND_{70}$, Sophie Masson)	:	hasPersonName)
($SB\text{-}IND_{71}$	:	*Person*),	(($SB\text{-}IND_{71}$, Viktor Vasnetsov)	:	hasPersonName)
($SB\text{-}IND_{72}$	:	*Person*),	(($SB\text{-}IND_{72}$, Walter Crane)	:	hasPersonName)
($SB\text{-}IND_{73}$	:	*Person*),	(($SB\text{-}IND_{73}$, William Makepeace Thackeray)	:	hasPersonName)
($SB\text{-}IND_{74}$	:	*Position*),	(($SB\text{-}IND_{74}$, Director of the Imperial Theatres)	:	hasPositionName)
($SB\text{-}IND_{75}$	:	*Position*),	(($SB\text{-}IND_{75}$, Princess)	:	hasPositionName)
($SB\text{-}IND_{76}$	:	*Position*),	(($SB\text{-}IND_{76}$,Queen)	:	hasPositionName)
($SB\text{-}IND_{77}$	:	*Position*),	(($SB\text{-}IND_{77}$, Sailor)	:	hasPositionName)
($SB\text{-}IND_{78}$	:	*Position*),	(($SB\text{-}IND_{78}$,chaplain)	:	hasPositionName)
($SB\text{-}IND_{79}$	:	*Position*),	(($SB\text{-}IND_{79}$, collector)	:	hasPositionName)
($SB\text{-}IND_{80}$	:	*Position*),	(($SB\text{-}IND_{80}$, king)	:	hasPositionName)
($SB\text{-}IND_{81}$	:	*Position*),	(($SB\text{-}IND_{81}$, king and queen)	:	hasPositionName)
($SB\text{-}IND_{82}$	:	*Position*),	(($SB\text{-}IND_{82}$, player)	:	hasPositionName)
($SB\text{-}IND_{83}$	:	*Position*),	(($SB\text{-}IND_{83}$, prince)	:	hasPositionName)
($SB\text{-}IND_{84}$	:	*Position*),	(($SB\text{-}IND_{84}$, prince and princess)	:	hasPositionName)
($SB\text{-}IND_{85}$	:	*Position*),	(($SB\text{-}IND_{85}$, queen)	:	hasPositionName)
($SB\text{-}IND_{86}$	:	*Position*),	(($SB\text{-}IND_{86}$, the king)	:	hasPositionName)
($SB\text{-}IND_{87}$	:	*Position*),	(($SB\text{-}IND_{87}$, secretary)	:	hasPositionName)
($SB\text{-}IND_{88}$	:	*Position*),	(($SB\text{-}IND_{88}$, skilled martial artist)	:	hasPositionName)
($SB\text{-}IND_{89}$	:	*Position*),	(($SB\text{-}IND_{89}$, the king)	:	hasPositionName)
($SB\text{-}IND_{90}$	:	*ProvinceOrState*),	(($SB\text{-}IND_{90}$, Aurora)	:	hasProvinceOrStateName)
($SB\text{-}IND_{91}$	:	*ProvinceOrState*),	(($SB\text{-}IND_{91}$, Calabria)	:	hasProvinceOrStateName)
($SB\text{-}IND_{92}$	:	*PublishedMedium*),	(($SB\text{-}IND_{92}$, L'Aurore)	:	hasPublishedMediumName)
($SB\text{-}IND_{93}$	:	*PublishedMedium*),	(($SB\text{-}IND_{93}$, Le Jour)	:	hasPublishedMediumName)
($SB\text{-}IND_{94}$	:	*TVShow*),	(($SB\text{-}IND_{94}$, Once Upon a Time)	:	hasTVShowName)
($SB\text{-}IND_{95}$	:	*TVShow*),	(($SB\text{-}IND_{95}$, Sleeping Beauty)	:	hasTVShowName)

Table A.3: Analysis Abox $\mathcal{A}_{d_9}$

Annotation of Document d_{10}

($SW\text{-}domain$	:	*FairyTale*),			
($SW\text{-}IND_{1}$	:	*Company*),	($SW\text{-}IND_{1}$, “ABC”)	:	hasCompanyName)
($SW\text{-}IND_{2}$	:	*Company*),	($SW\text{-}IND_{2}$, Disney Enterprises Inc.)	:	hasCompanyName)
($SW\text{-}IND_{3}$	:	*Company*),	($SW\text{-}IND_{3}$, Huntsman)	:	hasCompanyName)
($SW\text{-}IND_{4}$	:	*Company*),	($SW\text{-}IND_{4}$, Walt Disney)	:	hasCompanyName)
($SW\text{-}IND_{5}$	:	*Continent*),	($SW\text{-}IND_{5}$,Europe)	:	hasContinentName)
($SW\text{-}IND_{6}$	:	*Continent*),	($SW\text{-}IND_{6}$, North America)	:	hasContinentName)
($SW\text{-}IND_{7}$	:	*Country*),	($SW\text{-}IND_{7}$, Albania)	:	hasCountryName)
($SW\text{-}IND_{8}$	:	*Country*),	($SW\text{-}IND_{8}$, Armenia)	:	hasCountryName)
($SW\text{-}IND_{9}$	:	*Country*),	($SW\text{-}IND_{9}$, Germany)	:	hasCountryName)
($SW\text{-}IND_{10}$	:	*Holiday*),	($SW\text{-}IND_{10}$, peace day)	:	hasHolidayName)
($SW\text{-}IND_{11}$	:	*Movie*),	($SW\text{-}IND_{11}$, Blancanieves)	:	hasMovieName)
($SW\text{-}IND_{12}$	:	*Movie*),	($SW\text{-}IND_{12}$, Grimm)	:	hasMovieName)
($SW\text{-}IND_{13}$	:	*Movie*),	($SW\text{-}IND_{13}$, Into the Woods)	:	hasMovieName)
($SW\text{-}IND_{14}$	:	*Movie*),	($SW\text{-}IND_{14}$, Mirror Mirror)	:	hasMovieName)
($SW\text{-}IND_{15}$	:	*Movie*),	($SW\text{-}IND_{15}$, Rose Red)	:	hasMovieName)
($SW\text{-}IND_{16}$	:	*Movie*),	($SW\text{-}IND_{16}$, Snow White)	:	hasMovieName)
($SW\text{-}IND_{17}$	:	*Movie*),	($SW\text{-}IND_{17}$, Snow White and the Seven Dwarfs)	:	hasMovieName)
($SW\text{-}IND_{18}$	:	*Movie*),	($SW\text{-}IND_{18}$, Snow White: A Tale of Terror)	:	hasMovieName)
($SW\text{-}IND_{19}$	:	*Movie*),	($SW\text{-}IND_{19}$, Snow White: The Fairest of Them All)	:	hasMovieName)
($SW\text{-}IND_{20}$	:	*Movie*),	($SW\text{-}IND_{20}$, The Brothers Grimm)	:	hasMovieName)
($SW\text{-}IND_{21}$	:	*Organization*),	($SW\text{-}IND_{21}$, US Patent and Trademark Office)	:	hasOrganizationName)
($SW\text{-}IND_{22}$	:	*Organization*),	($SW\text{-}IND_{22}$, group of seven)	:	hasOrganizationName)

$(SW\text{-}IND_{23}$	:	$Person$),	$(SW\text{-}IND_{23}$, Adolph Zukor)	:	hasPersonName)
$(SW\text{-}IND_{24}$	:	$Person$),	$(SW\text{-}IND_{24}$, Alexander Pushkin)	:	hasPersonName)
$(SW\text{-}IND_{25}$	:	$Person$),	$(SW\text{-}IND_{25}$, Alexander Zick Folk tale)	:	hasPersonName)
$(SW\text{-}IND_{26}$	:	$Person$),	$(SW\text{-}IND_{26}$, Andrew Alcott)	:	hasPersonName)
$(SW\text{-}IND_{27}$	:	$Person$),	$(SW\text{-}IND_{27}$, Bill Willingham)	:	hasPersonName)
$(SW\text{-}IND_{28}$	:	$Person$),	$(SW\text{-}IND_{28}$, Brangomar)	:	hasPersonName)
$(SW\text{-}IND_{29}$	:	$Person$),	$(SW\text{-}IND_{29}$, Brighton)	:	hasPersonName)
$(SW\text{-}IND_{30}$	:	$Person$),	$(SW\text{-}IND_{30}$, Charlize Theron)	:	hasPersonName)
$(SW\text{-}IND_{31}$	:	$Person$),	$(SW\text{-}IND_{31}$, Chris Hemsworth)	:	hasPersonName)
$(SW\text{-}IND_{32}$	:	$Person$),	$(SW\text{-}IND_{32}$, Clementianna)	:	hasPersonName)
$(SW\text{-}IND_{33}$	:	$Person$),	$(SW\text{-}IND_{33}$, Creighton Hale)	:	hasPersonName)
$(SW\text{-}IND_{34}$	:	$Person$),	$(SW\text{-}IND_{34}$, Daniel Frohman)	:	hasPersonName)
$(SW\text{-}IND_{35}$	:	$Person$),	$(SW\text{-}IND_{35}$, Diana Rigg)	:	hasPersonName)
$(SW\text{-}IND_{36}$	:	$Person$),	$(SW\text{-}IND_{36}$, Dorothy Cumming)	:	hasPersonName)
$(SW\text{-}IND_{37}$	:	$Person$),	$(SW\text{-}IND_{37}$, Elizabeth McGovern)	:	hasPersonName)
$(SW\text{-}IND_{38}$	:	$Person$),	$(SW\text{-}IND_{39}$, Evil Queen)	:	hasPersonName)
$(SW\text{-}IND_{40}$	:	$Person$),	$(SW\text{-}IND_{40}$, Florimond)	:	hasPersonName)
$(SW\text{-}IND_{41}$	:	$Person$),	$(SW\text{-}IND_{41}$, Ginnifer Goodwin)	:	hasPersonName)
$(SW\text{-}IND_{42}$	:	$Person$),	$(SW\text{-}IND_{42}$, Johann Georg von Hahn)	:	hasPersonName)
$(SW\text{-}IND_{43}$	:	$Person$),	$(SW\text{-}IND_{43}$, Julia Roberts)	:	hasPersonName)
$(SW\text{-}IND_{44}$	:	$Person$),	$(SW\text{-}IND_{44}$, Kristen Stewart)	:	hasPersonName)
$(SW\text{-}IND_{45}$	:	$Person$),	$(SW\text{-}IND_{45}$, Kristin Kreuk)	:	hasPersonName)
$(SW\text{-}IND_{46}$	:	$Person$),	$(SW\text{-}IND_{46}$, Lily Collins)	:	hasPersonName)
$(SW\text{-}IND_{47}$	:	$Person$),	$(SW\text{-}IND_{47}$, Marguerite Clark)	:	hasPersonName)
$(SW\text{-}IND_{48}$	:	$Person$),	$(SW\text{-}IND_{48}$, Mary Jane)	:	hasPersonName)
$(SW\text{-}IND_{49}$	:	$Person$),	$(SW\text{-}IND_{49}$, Miranda Richardson)	:	hasPersonName)
$(SW\text{-}IND_{50}$	:	$Person$),	$(SW\text{-}IND_{50}$, Mirror Queen)	:	hasPersonName)
$(SW\text{-}IND_{51}$	:	$Person$),	$(SW\text{-}IND_{51}$, Monica Keena)	:	hasPersonName)
$(SW\text{-}IND_{52}$	:	$Person$),	$(SW\text{-}IND_{52}$, Nagamati)	:	hasPersonName)
$(SW\text{-}IND_{53}$	:	$Person$),	$(SW\text{-}IND_{53}$, Nathan Lane)	:	hasPersonName)
$(SW\text{-}IND_{54}$	:	$Person$),	$(SW\text{-}IND_{54}$, Nicola Stapleton)	:	hasPersonName)
$(SW\text{-}IND_{55}$	:	$Person$),	$(SW\text{-}IND_{55}$, Prince)	:	hasPersonName)
$(SW\text{-}IND_{56}$	:	$Person$),	$(SW\text{-}IND_{56}$, Ravenna)	:	hasPersonName)
$(SW\text{-}IND_{57}$	:	$Person$),	$(SW\text{-}IND_{57}$, Rupert Sanders)	:	hasPersonName)
$(SW\text{-}IND_{58}$	:	$Person$),	$(SW\text{-}IND_{58}$, Sam Claflin)	:	hasPersonName)
$(SW\text{-}IND_{59}$	:	$Person$),	$(SW\text{-}IND_{59}$, Sam Neill)	:	hasPersonName)
$(SW\text{-}IND_{60}$	:	$Person$),	$(SW\text{-}IND_{60}$, Sarah Patterson)	:	hasPersonName)
$(SW\text{-}IND_{61}$	:	$Person$),	$(SW\text{-}IND_{61}$, Sigourney Weaver)	:	hasPersonName)
$(SW\text{-}IND_{62}$	:	$Person$),	$(SW\text{-}IND_{62}$, Stephen Sondheim)	:	hasPersonName)
$(SW\text{-}IND_{63}$	:	$Person$),	$(SW\text{-}IND_{63}$, Vanessa Redgrave)	:	hasPersonName)
$(SW\text{-}IND_{64}$	:	$Person$),	$(SW\text{-}IND_{64}$, Snow White)	:	hasPersonName)
$(SW\text{-}IND_{65}$	:	$Person$),	$(SW\text{-}IND_{65}$, William)	:	hasPersonName)
$(SW\text{-}IND_{66}$	:	$Person$),	$(SW\text{-}IND_{66}$, Zwerge)	:	hasPersonName)
$(SW\text{-}IND_{67}$	:	$Position$),	$(SW\text{-}IND_{67}$, King)	:	hasPositionName)
$(SW\text{-}IND_{68}$	:	$Position$),	$(SW\text{-}IND_{68}$, Prince)	:	hasPositionName)
$(SW\text{-}IND_{69}$	:	$Position$),	$(SW\text{-}IND_{69}$, Princess)	:	hasPositionName)
$(SW\text{-}IND_{70}$	:	$Position$),	$(SW\text{-}IND_{70}$, farmer)	:	hasPositionName)
$(SW\text{-}IND_{71}$	:	$Position$),	$(SW\text{-}IND_{71}$, queen)	:	hasPositionName)
$(SW\text{-}IND_{72}$	:	$Position$),	$(SW\text{-}IND_{72}$, queen and king)	:	hasPositionName)
$(SW\text{-}IND_{73}$	:	$Position$),	$(SW\text{-}IND_{73}$, teacher)	:	hasPositionName)
$(SW\text{-}IND_{74}$	:	$TVShow$),	$(SW\text{-}IND_{74}$, Faerie Tale Theatre)	:	hasTVShowName)
$(SW\text{-}IND_{75}$	:	$TVShow$),	$(SW\text{-}IND_{75}$, Once Upon a Time)	:	hasTVShowName)
$(SW\text{-}IND_{76}$	:	$TVShow$),	$(SW\text{-}IND_{76}$, The 10th Kingdom)	:	hasTVShowName)

Table A.4: Analysis Abox $\mathcal{A}_{d_{10}}$