GrOnto: a GRanular ONTOlogy for Diversifying Search Results

Silvia Calegari
Gabriella Pasi
University of Milano-Bicocca
V.le Sarca 336/14, 20126
Milano, Italy
{calegari,pasi}@disco.unimib.it

ABSTRACT

Results diversification is an approach used in literature to cover the possible interpretations of the results produced by query evaluation. For diversifying search results we propose the GrOnto model. This model is based on a normalized granular view of an ontology: GrOnto allows to associate each result with the suited topical granules in order to categorize it based on the granular information.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval- *information filtering, search process*

1. INTRODUCTION

In last years, Web search engines have become the de-facto access point to the information available on the Internet. Usually people specify their information needs by writing queries with a limited number of terms (usually $2-3$ terms per query). However, short queries are very difficult to disambiguate: in fact a term may have several interpretations. One of the problems related to term disambiguation is how to diversify results produced as an answer to an ambiguous query. An interesting research topic that in recent years has attracted several researchers is results diversification. The focus is on how to produce a set of diversified results that cover the different possible interpretations of the query. The importance of result diversification has been recognized as a very important topic in Information Retrieval; the basic idea is that "*the relevance of a set of documents depends not only on the individual relevance of its members, but also on how they relate to one another*"[3]. The key aspect is that the relevance of a document has to consider also the semantics expressed by the terms it contains."*The focus is on how to diversify search results making explicit use of knowledge about the topics the query or the documents may refer to*" [1].

In a recent research work, a taxonomy of information is used to model the user's request [1]. The idea is to assign both query and documents to one or more categories of the taxonomy. The taxonomy adopted is the one provided by the ODP [1] ontology. Furthermore, it is assumed that *usage statistics have been collected on the distribution of user intents over the categories* ([6]). The aim of this approach is to minimize the risk of user dissatisfaction by computing a *quality value* for each document retrieved in response to a query as a combination of relevance and diversity.

In this paper a method for diversifying the results produced in response to a query is proposed. We do not use a statistical approach in order to diversify the results, but our method makes use of a semantic support offered by a granular view of an ontology [2] to the aim of producing a granular taxonomy of the results. By this method the information is classified at different topical levels (from a general topic to a specific topic).
In a granular ontology the concepts and instances are classified into granules. A granule is a chunk of knowledge made of different objects "*drawn together by indistinguishability, similarity, proximity or functionality*"[12]. A level is just the collection of granules of similar nature, and a granular information is a pyramidal information structure with different levels of clarifications.
The paper is organized as follows. In Section 2 an overview of the use of ontologies in Information Retrieval is presented. In Section 3 the definition of a normalized granular view of an ontology is reported. The approach proposed in this work, named GrOnto, for diversifying search results is defined in Section 4. At the end, in Section 5 some conclusions and future works are stated.

2. THE USE OF ONTOLOGIES IN INFORMATION RETRIEVAL

In the last decades ontologies have been used in different areas of research in Computer Science, among which Information Retrieval where they have been involved into several applications to different aims. For example, ontologies have been used: in distributed environments, for reranking the results to better satisfy the user's needs, to provide conceptual indexing and to disambiguate user's query. In distributed environment, significant works are SemreX [7] and Semantic Link Network (SLN)[13]. SemreX is a recent project that implements a multi-layer overlay network to map semantically correlated documents to clustered groups of neighbors. This semantic mapping is obtained by considering the ACM Topic Ontology. In SLN, an ontology has

[1]ODP: Open Directory Project, (http://dmoz.org)

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

been built as a self-organized semantic data model by defining semantic nodes, semantic links among nodes, and a set of relational reasoning rules; where each node identifies a resource.
In order to re-rank the results obtained after a search on the Web, generally, a user's profile is used. In the literature different strategies have been defined in order to build a user's profile by adopting the semantic support of an ontology. For example in [4] a user profile is built by considering past queries, and it is represented as a weighted graph by extracting the related terms from the ODP ontology.
In the conceptual indexing field of research, WordNet[2] synsets are used as terms for the representation of the documents. The concept detection phase consists in extracting concepts from documents that correspond to synsets in WordNet. In [8] the authors proposed some procedures to identify the correct sense of a word.
In this paper we are interested in the last field of research where the problem of disambiguation of the query is taken into account. Short queries are very difficult to disambiguate. Two main problems may arise: word synonymy (i.e., two words with the same meaning), and word polysemy (i.e., one word with multiple meanings). In the literature several strategies have been proposed in order to find a solution to this problem. Also ontologies have been involved in this field with the goal to provide a semantic support for reducing the ambiguity of the query. A way is to analyse the structure of the ontology to expand the terms written into the query with new meanings terms. The use of ontology reduces the possible (mis)interpretation of a query, but it needs to tune a query term to the right level in the hierarchy. Not only the IS-A relationship is used to discover the suited words [11], but also other important relationships such as, synonymy, meronymy and hypernyms are taken into account. For example in [9] the relationships considered are: *hyperonymy* and *synset*. For each term written in the query, a set of its synsets in WordNet is identified.

As reported in the Introduction of this paper, the results diversification is another strategy that can be adopted to solve the problem of ambiguous queries. We are interested in the situation where there is the necessity to individuate the different interpretations of a user's query. The focus is to produce a set of diversified results that cover at best these interpretations. One of pioneers works on diversification is that of Carbonell and Goldstein [3]. In their work, the diversification is obtained through the use of two similarity functions: one for measuring the similarity of the documents, and the other one for measuring the similarity between each document and a query. In more recent works a new approach has been explored to categorize both queries and documents by the use of a taxonomy [1, 14]. In these papers the taxonomy adopted is the one of the ODP ontology. The taxonomy is set by the IS-A relationship among categories; in fact in this context each concept of the ODP ontology represents a specific category.
In our paper we propose a method to diversify search results with the adoption of a new granular view of an ontology. Whereas in the previous works ([1, 14]) the taxonomy has been used only as a vocabulary for individuating the categories for queries and documents, now we consider an innovative ontology framework with a semantic expressiveness (i.e., instances and their properties) richer than the ODP ontology.

[2]http://wordnet.princeton.edu/

3. GRANULAR VIEW OF AN ONTOLOGY

This proposed method is based on the concept of a granular view (or granular perspective) of an ontology which has been defined in [2]. Given a domain ontology, the idea is to analyse the instances and their properties in order to discover new semantic associations among them. These semantic associations can be defined with the application of a rough methodology. The objective is to re-organize the ontology in a new taxonomy obtained after the analysis of the properties values assigned to the instances.
The rough structure used is known as Information Table [10]. For a domain ontology, an Information Table is induced as the structure:

$$\langle I, P, Val(I), F \rangle$$

where I is the set of the instances, P is the set of the properties, $Val(I)$ is the set of all the values assumed by the properties P, and F is the function that assigns to a pair (i, p) the value assumed by the instance $i \in I$ on the property $p \in P$. Thus, we can say that two instances are similar if they have the same values only for some properties. Formally, let $D \subseteq P$, then given two instances $i_1, i_2 \in I$, i_1 is similar to i_2 with respect to D and ϵ, with $\epsilon \in [0, 1]$, iff

$$\frac{|\{d_j \in D : F(i_1, d_j) = F(i_2, d_j)\}|}{|D|} \geq \epsilon \qquad (1)$$

This relation says that two instances are similar if they have at least $\epsilon|D|$ properties with the same value. For example, if we consider a Wine Ontology then a possible set of properties is $P := \{Location, Color, Sugar, Flavor, Body\}$. D is a subset of P defined as $D := \{Sugar, Flavor, Body\}$. In this case two instances belong to the same granule if they have at least $|(D-1)|$ properties with the same value, i.e. $\epsilon := \frac{|(D-1)|}{|D|} := \frac{2}{3}$. For example, *Longridge Merlot* and *Marietta Zinfandel* belong to the same granule by having two properties with the same value, i.e. $(flavor == moderate)$ and $(sugar == dry)$.

In [2] the instances are classified into granules at a different level of clarification. A key aspect is how to choose the granular levels from the non-granular ontology. The idea is to cluster the instances into granules by considering their similarity, i.e. by analysing the values of their properties (see Equation 1).

The granular view of an ontology is defined by following 3 steps. In order to clarify the construction of the new ontology, we refer to a very simple example. In this example, let us consider a small Wine Ontology which has 4 instances, and the set P of properties previously defined.
First step: definition of the tabular version of the ontology. In this table the rows are the instances and the columns are all the properties defined in the ontology. The selected instances and properties are the ones defined only by the IS-A relationships of the ontology domain. Table 1 reports the instances and the properties with their values of the small Wine Ontology analysed in this work.
Second step: It consists in the definition of the granular levels. As previously stated the granular levels have been chosen by analysing the properties values of the instances.

Table 1: A tabular version for the small Wine Ontology

Instances	Color	Sugar	Flavor	Body	Location
`Longridge Merlot`	Red	Dry	Moderate	Light	$Undefined$
`Marietta Zinfandel`	Red	Dry	Moderate	Medium	$Undefined$
`Lane Tanner Pinot Noir`	Red	Dry	Delicate	Light	$Undefined$
`Chateau-D-Ychem`	$Undefined$	$Undefined$	$Undefined$	$Undefined$	Bordeaux region

The tabular representation is used as support for this step. Thus, from the set of properties P two disjoint sets of granules are induced: $D_1 := \{Color, Flavor, Body, Sugar\}$ and $D_2 := \{Location\}$. Only *Location* belongs to the first level with the instance $Chateau-D-Ychem$ at the second granular level. Whereas for D_1, the choice of the first granular level has to be made among the properties that belong to D_1. Also in this case we have to analyze the properties values assumed by the set of instances, and we can observe that the identification of the first granular level can be made arbitrarily between *Color* and *Sugar* since they assume the same values for all their instances. For this ontology, without loss of generality, we can consider *Color* at the first granular level, and for the next level the similarity relation (i.e., Equation 1) to the D_1 set (without the property *Color*) can be applied. In this illustrative example $\epsilon := \frac{2}{3}$, that is, two instances belong to the same granule if they have at least two out of three properties with the same value. Figure 1 depicts the granular classification obtained where the circles are the properties values and the squares are the instances.

The **third step** is to solve the problem of redundancy of

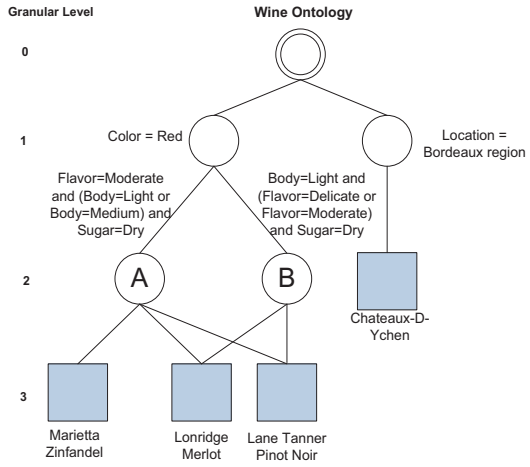

Figure 1: A granular view of a small Wine Ontology after the application of the rough methodology.

the information. Let us consider two granules G_i and G_j at the same granular level, we have that G_i is redundant with respect to G_j iff $G_j \supseteq G_i$. In [2] a normalisation process has been defined in order to obtain a normal form of the granular perspective. For example, if we examine the same example of Figure 1, we can observe that G_A and G_B belong to the same granular level, and that $G_A \supseteq G_B$. Indeed, the instances *Lonridge Merlot* and *Lane Tanner Pinot Noir* are completely included into G_B but they belong to G_A. In this normalisation process the granular subclass G_B inherits all the common instances from the granular superclass G_A (see Figure 2).

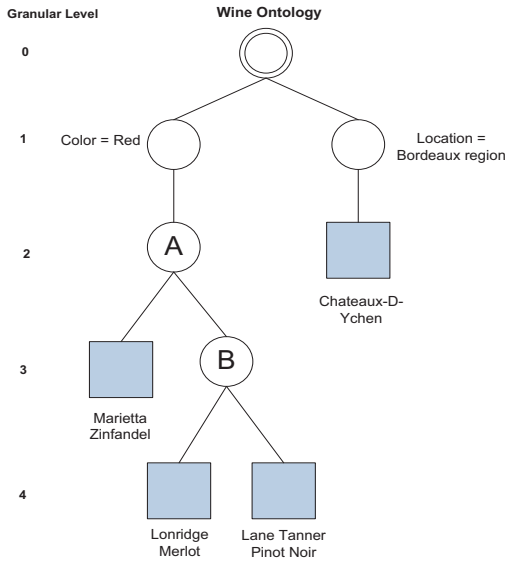

Figure 2: The granular view of the small Wine Ontology after the application of the normalisation process.

4. THE PROPOSED MODEL

When using a search engine a user formulates a query in order to retrieve the documents relevant to her/his information needs. In most cases the user writes short queries that are difficult to disambiguate. In fact, in several user's queries a query term could be interpreted with different meanings. We propose a solution to diversify search results that aims to increase the effectiveness of the system by reducing the ambiguity in the interpretation of results. As proposed in [1] we adopt a taxonomy of information where both queries and results may belong to more than one category. In particular we use the taxonomy corresponding to a normalized granular view of an ontology (see Section 3). The idea is to associate each result with the suited topical granules.

Generally, in search engines the evaluation of a user's query produces an ordered list of results. For diversifying search results the GrOnto model (see Figure 3) takes in input a ranked *list of results*, and the *granular ontology* to categorize each result. In other words, the normalized granular view of the ontology is used to apply a filtering on the search results. As reported in Section 1, in a *granular* ontology the granules are organized at different levels of clarifications. Thus the categorization of each result is performed by locating in the ontology the right granules with which it may be associated. Figure 4 shows the general structure of the approach where the list of results (left-hand side of Figure 4) is re-organized by the filtering strategy (right-hand side of Figure 4) based on the *granular* ontology structure. By applying the catego-

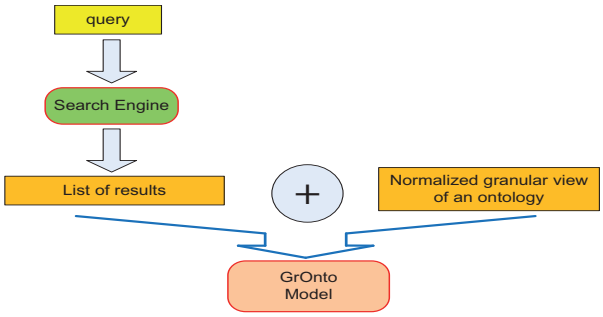

Figure 3: A simple schema of the GrOnto model.

rization process (explained here below), we obtain a representation of the results which reflects the classification into topics corresponding to the granular levels of the adopted ontology. Each retrieved document is associated with one

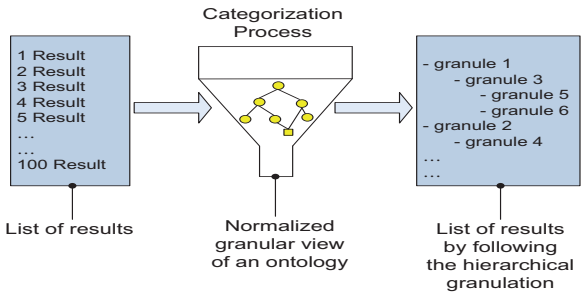

Figure 4: A Web search after the application of the GrOnto model.

or more granules of the ontology by a procedure explained here below.

As an example, let us consider the same vocabulary and structure of the Wine Ontology described in Section 3. The related set of concepts is $O := \{Red, Bordeaux\,region, Chateau-D-Ychen, MariettaZinfandel, Lonridge\,Merlot, Lane\,Tanner\,Pinot\,Noir\}$. During a search session a user is interested in finding, for instance, information about *red wines* and she/he writes the following short query *q:="red wines in France"*, and a list of results is displayed. The association of each result with granules of the *granular* ontology is obtained in two steps. Here below the process undertaken to categorize a search result is explained. We present these two steps in order to categorize the first result, obviously the same procedure is applied to the other search results.

Step 1: "Formal representation of each result". In order to formally represent the content of a result R_i proposed in response to a query, we assume that results are described by *Title* and *Snippet*. The $i-th$ result R_i is then associated with a set of terms, Res_i, extracted from the textual information, i.e. $Res_i := Title_i \cup Snippet_i$ where $Title_i$ and $Snippet_i$ are sets of terms included into the vocabulary of the *granular* ontology.

Thus, by analysing the first result R_1, we have: *Title:="Wines of France-A guide to French wines"* and *Snippet:="Discover the wines of France, their varieties, history and regions;... Lane Tanner Pinot Noir is a very famous red wine produced in..."*. From these two short texts, by considering the set O, we obtain that $Res_1 := \{Lane\,Tanner\,Pinot\,Noir, Red\}$, i.e. $Title_1 := \emptyset := Title \cap O$ and $Snippet_1 := \{Lane\,Tanner\,Pinot\,Noir, Red\} := Snippet \cap O$.

Step 2: "Association of each result R_i with granules of the granular tree". The output of Step 1 is a set of terms of the vocabulary O, named Res_i, for each retrieved document R_i. An element of Res_i is a granule of the ontology, and to this granule we can associate the $i-th$ result. Thus, for each granule the following structure: $< Results_j, card_{TOT_j} >$ is defined, where $Results_j$ is the set of the search result associated with the $j-th$ granule, i.e. $Results_j := \{R_i | granule_j \in Res_i\}$, and $card_{TOT_j}$ is the cardinality of all the results associated with the $j-th$ granule. This means that $card_{TOT_j} := |Results_j \cup (\bigcup_{child=0}^{n} Results_{child})|$ i.e., the cardinality of all the results individuated with the granule $j-th$ and the cardinality of the results associated with all its n sub-granules (children nodes).

By considering the same example of Step 1, we have that the first result R_1 has been formally represented as $Res_1 := \{Lane\,Tanner\,Pinot\,Noir, Red\}$ so that, the selected granules are *Lane Tanner Pinot Noir* and *Red*. Figure 5 depicts the situation after the application of Step 2 where the structure assigned with $granule_1$ is $< Results_1 := \{R_1\}, 1 >$, whereas for $granule_8$ is $< Results_8 := \{R_1\}, 1 >$. Thus, we have that the first result R_1 has been categorized with two topics (granules) at a different level of clarification.

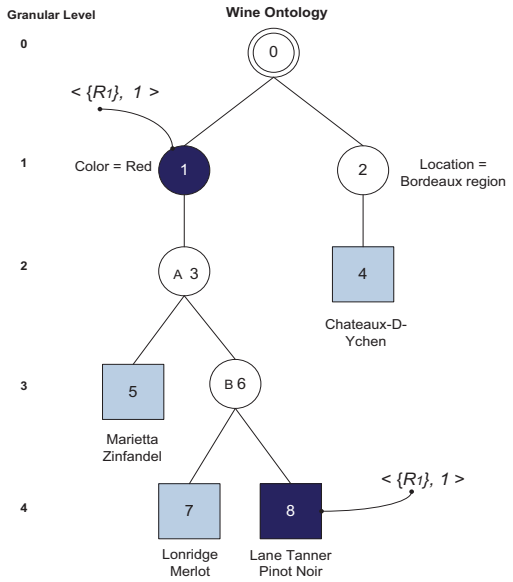

Figure 5: Example of the structure assigned to each granule identified with a result.

GrOnto on the Web.

Figure 6 depicts a prototype interface for the GrOntoS system. We have taken inspiration from *Clusty*[3] where the web-page structure is split into three parts: 1) a text area where the user can formulate her/his request by using the Yahoo! Search engine, 2) a profile used to visualize the portion of the normalized granular view of the ontology involved from the specific query, and 3) a web-page area devoted to the visualization of the results. In particular only the results categorized with a granule of the ontology are displayed

[3](http://clusty.com/)

one by one. Figure 6 reports a simple example where the small Wine Ontology of Section 3 is used to classify *ALL* the results obtained, for example, after the evaluation of the *q:=red wines in France*. A user can use the portion of the granular ontology in order to navigate the results by considering the categorization provided by the levels granular. In fact by clicking on an item of the portion of the granular ontology, all its results will be visualised. Furthermore, each item is enriched with the cardinality of the results associated with its topic, in this way the user is directed towards the category more numerous.

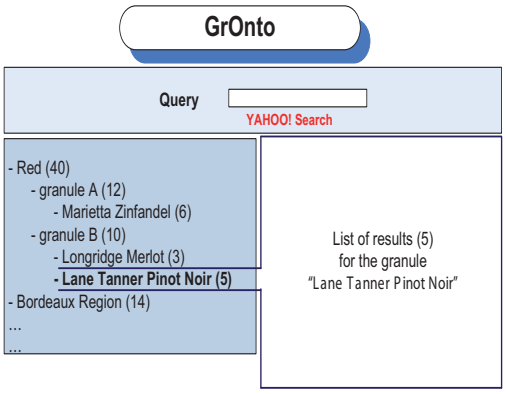

Figure 6: The interface model of the GrOnto model.

5. CONCLUSIONS

In this paper we have studied the problem of diversification of search results to disambiguate the user's query in a given domain of knowledge represented by a granular ontology. We have proposed a model, named Gronto, based on a semantic support for associating search result with one or more categories. A normalized granular view of an ontology is the semantic framework adopted in order to cover all the possibles meanings of a result. Generally, after the evaluation of a user's query an ordered list of results is obtained. GrOnto takes in input this list and the granular ontology, and thanks to the adoption of a filtering strategy a taxonomic organization of the results is achieved.

We are implementing the GrOnto model through a simple web service by adopting the representational state transfer (REST) paradigm [5].

The prosecution of this research activity will address the problem of applying the GrOnto approach to personalized ontologies, where the user interests will be represented by means of a granular ontology. To this aim we are also investigating the problem of defining personalized granular ontologies.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, NY, USA, 2009. ACM.

[2] S. Calegari and D. Ciucci. Granular computing applied to ontologies. *International Journal of Approximate Reasoning*, 2009. In printing.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *In Research and Development in Information Retrieval*, pages 335–336, 1998.

[4] M. Daoud, L. Tamine-Lechani, M. Boughanem, and B. Chebaro. A session based personalized search using an ontological user profile. In S. Y. Shin and S. Ossowski, editors, *SAC*, pages 1732–1736. ACM, 2009.

[5] R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. In *ICSE '00: Proceedings of the 22nd international conference on Software engineering*, pages 407–416, New York, NY, USA, 2000. ACM.

[6] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 61–70, New York, NY, USA, 2008. ACM.

[7] H. Jin and H. Chen. Semrex: Efficient search in a semantic overlay for literature retrieval. *Future Generation Computer System*, 24(6):475–488, 2008.

[8] R. Mihalcea and D. I. Moldovan. Semantic indexing using wordnet senses. In *In Proceedings of ACL Workshop on IR & NLP*, pages 35–45, 2000.

[9] R. Navigli and P. Velardi. An analysis of ontology-based query expansion strategies. In *Workshop on Adaptive Text Extraction and Mining, (Cavtat Dubrovnik, Croatia, Sept 23)*, 2003.

[10] Z. Pawlak. Information systems - theoretical foundations. *Information Systems*, 6:205–218, 1981.

[11] E. M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[12] L. Zadeh. Is there a need for fuzzy logic? *Information Sciences*, 178:2751–2779, 2008.

[13] H. Zhuge. Communities and emerging semantics in semantic link network: Discovery and learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):785–799, 2009.

[14] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, New York, NY, USA, 2005. ACM.