Extending an on-line information site with accurate domain-dependent extracts from the World Wide Web[⋆]

Enrique Alfonseca and Pilar Rodríguez

Computer Science Department, Universidad Autonoma de Madrid,
Carretera de Colmenar Viejo, km. 14,5,
28043 Madrid, Spain
{Enrique.Alfonseca,Pilar.Rodriguez}@ii.uam.es

Abstract. This paper describes a new procedure that has been developed for extending an existing on-line information system about *The Voyages of the Beagle* with information collected automatically from Internet. A Term Identification procedure finds relevant terms in the document; and the algorithm uses conventional search engines (such as `Google`) to look for pages about those terms. Next, a sequence of filters rule out all the information considered irrelevant, and the remaining data is put together in "*summary pages*" available to the students. Our experiments so far have attained very good results, and in a form that was sent to several users of the on-line site they all showed much excitement about the tool.

Keywords: Knowledge Acquisition, Internet search, Query formulation, Filters

1 Introduction

In the last years, the Internet has consolidated as a useful source of information of any kind. It provides a high versatility for text and multimedia presentations and, at the same time, it is easy for an author to modify, add and remove documents. However, the very same characteristics that make it useful and popular can hinder the search and analysis of relevant information. The amount of information is so overwhelming that, when performing a search, many of the results provided by search engines are incomplete, outdated or irrelevant for our purposes.

When a search is performed on a vast repository such as the Internet, it is necessary to filter accurately the retrieved texts in order to guarantee that the information is really relevant. Most likely, this filtering will be performed by a human that reviews the web pages by hand. For example, by looking on Internet for information about the book *Don Quixote*, we are able to find data about films, TV series, restaurants, exhibits, schools, reading groups, etc. If the collection of data is to be automatic, there have to be means in which to filter out all this unwanted information.

We have applied an automatic algorithm for extending an on-line site about Darwin's *The Voyages of the Beagle* that is accessible for the students in the Universidad Autónoma de Madrid. Additional information has been collected about all the locations and some other terms cited in Darwin's text, and the results are encouraging. A Term

[⋆] This work has been sponsored by CICYT, project number TIC2001-0685-C02-01.

Identification procedure was used to suggest possible relevant terms from the text. Also, it is possible for the students to provide any other term cited in the text. For instance, a student may want to know additional information about a city visited by Darwin (e.g. Valparaiso) or about an animal or a plant. For each term, the system generates one summary page, by filtering out all the downloaded information that was judged irrelevant.

To do this, several filters are applied to the set of retrieved pages. The main stress has been placed on precision, rather than on recall. We have observed that the web contains some amount of redundancy, and thence the loss of a web page does not necessarily mean that the information it contains will be lost, because other page might contain it as well. Therefore, it has been considered more important that the information that is finally retrieved is relevant with respect to the user query. Natural Language Processing techniques are used to some extent during the filtering step.

1.1 Related work

The problem of finding information that is relevant to a user is the main topic of Information Retrieval [Baeza-Yates and Ribeiro-Neto, 1999]. The user need is usually stated with a query, either as a natural language question, a set of keywords, an expression with logical operators, or a set of documents that the user finds interesting. Research in this area was encouraged by the Text REtrieval Conferences (TRECs) [Voorhees and Harman, 2001], organised now for more than ten years by the National Institute of Standards and Technology (NIST), a competition in which the participant systems compete in locating relevant information.

From the semantic web community there are several initiatives aiming at annotating web pages with semantic labels, such as whether a term refers to a person or a location. The annotation of the web pages with semantics can be done with automatic approaches, such as Information Retrieval or Information Extraction procedures [Craven et al., 1999].

Some approaches to IR look for relevant information by reducing the search space, searching only in restricted portions of the World Wide Web. McCallum et al. [1999] describes a way in which a domain-specific search engine can be constructed from a set of initial documents and reinforced learning; and Rennie and McCallum [1999] also apply reinforced learning for crawling the Internet in search of domain-specific information. Diligenti et al. [2000] describes a focused crawler that looks for relevant web pages by modelling the hypertext contexts in which these might appear.

Query expansion using thesauri is one of the mechanisms used in IR for finding those documents that do not contain the keywords written by the user, but which contain synonyms of them. It can also be used in order to guide the search toward relevant data [Lawrence, 2000] [Boley et al., 1999].

2 Generation of summaries

Figure 1 describes the main architecture of the system. The input consists of two materials: one or several documents, which might be electronic text without any annotation, HTML pages, or XML documents; and a list of terms, for each of which the user needs

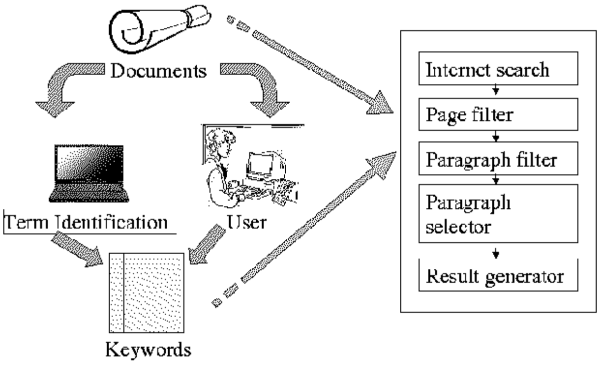


Fig. 1. General architecture of the system.

to collect information from the Internet. The keywords can be directly provided by the students, or it is also possible to use a Term Identification procedure that looks for relevant terms in the source document.

The purpose of the system is to produce, automatically, a summary page for each of the terms, with information collected automatically from the Internet. The documents provided must contain the terms, because they will be used during the filtering step, in order to discover which are the web pages in which a term is used with the same meaning than in the document. The idea is that the students may extend the information from the on-line site about *The Voyages of the Beagle* with additional information about the places he visited or the animals he studied.

The approach taken for generating each summary page is the following:

1. Firstly, a search is performed with the `Google` Internet search engine, using special keywords in order to increase the probability that the retrieved pages are relevant to the topic in question.
2. The format of the retrieved pages is changed into XML, and a standard linguistic processing (as described below) is performed on them.
3. A pipeline with two filters is applied. The first filter selects the relevant documents, and the second one selects the paragraphs in those documents. All the final information is put together in a single XML page.

2.1 Pre-processing and Terms Identification

The original document (*The Voyages of the Beagle* in our case) was pre-processed with the following tools:

- A tokeniser and a sentence-splitter written with regular expressions, in flex.
- The TnT part-of-speech tagger [Brants, 2000].
- A stemmer written in flex based on the LaSIE stemmer [Gaizauskas et al., 1995].
- Three chunkers in C++ and Java, to detect complex quantifiers, base noun phrases, and complex verbs, using transformation lists [Ramshaw and Marcus, 1995].

– A subject-verb and verb-object detector, written by us in Java *ad hoc*.
– A Named-Entity identifier that classifies some terms, such as locations, people, animals, artifacts or bodies of water [Alfonseca and Manandhar, 2002].

The user can specify which are the terms for which the summary pages are to be generated. However, in the case of our on-line Information system, we have used a simple automatic Term Identification module. It collects all the common words from the document that do not appear in a general-purpose dictionary, and every sequence of proper nouns, and counts their frequency of appearance. A term is kept if its frequency is above a given threshold (50).

2.2 Internet Search

For each term, a search is performed on Internet to find relevant information about it. The query to the search engine must contain the term words, as it is necessary some additional information in order to guide the search; otherwise, large amounts of irrelevant data would be returned.

As said before, several locations, people, dates, and other kinds of entities have been identified and labelled. A possible way to guide the search would be to include all the entities in the context of the term of interest. However, when this procedure was tried, the result was that many of the retrieved web pages contained fragments of *The Voyages of the Beagle*, as there are several copies of the book on-line.

In the final settings, the total number of entities that are added to the query is limited up to a maximum; furthermore, these entities are ordered depending on their type, and those that represent locations or bodies of water are given priority over the rest. This is because empirically it was observed that, when looking for information about either a person, a location, an animal or an artifact, the locations that appear in the context of the term are useful to find relevant pages among those returned by the search engine; while those that represent people, for example, tend to retrieve several copies of the original document on-line, and no additional information.

For illustration, the query for *Valparaiso* was generated in the following way:

1. Look for all the appearances of *Valparaiso* in the original document.
2. Get all the paragraphs where it appears.
3. Collect, from those paragraphs, the entities that are labelled as locations.
4. Add them, up to a maximum, to the query, as optional keywords.

The query produced is the following:

> “Valparaiso” (“Chile” OR “Chonos Archipelago” OR “S. Carlos” OR “Coquimbo” OR “Talcahuano” OR “Copiapo” OR “Concepcion” OR “Callao” OR “Cordillera” OR “Bahia Blanca” OR “Banda” OR “Peru” OR “Lima”)

Using this query, up to one hundred URLs are retrieved, and the documents are downloaded, translated into XML, and processed with the same linguistic tools, in order to obtain a shallow syntactic analysis of the sentences.

2.3 Page filter

Once the web pages have been retrieved and processed, the first filtering step eliminates the complete web pages that are judged irrelevant. It is necessary to make sure that the documents really refer to the term used in the right sense. For example, there were two documents concerning a different city called *Valparaiso*, which is located in Indiana, in the U.S., and other documents referring to *Valparaiso* as the name of a company, or included in a more complex name (e.g. *University of Valparaiso* or *Port Valparaiso Authorities*). To do the filtering, the web pages are compared one by one with the original document, using the procedure described in this section.

In order to discover which of the citations of the concept of interest are used with the requested meaning, we also used contextual and co-occurrence information, and the lexical semantic network WordNet [Miller, 1995].

WordNet is a semantic network of concepts, in which words are grouped in sets of synonyms that represent the same concept, called *synonym sets* or *synsets*. The version we used, 1.7, has 109,376 different synsets that cover nouns, verbs, adjectives and adverbs. These synsets are related to each other through several relationships. In particular, every synset in WordNet contains, apart from synonym words, a brief definition with its meaning. We can assume that the words that appear in the definition of a concept are related to it, so it is possible to define an equivalence relationship $\mathcal{G}$ between concepts, called *gloss*, in the following way: two concepts c and d are considered to be related in if either of the following holds: c appears in the definition of d, d appears in the definition of c, or there is a concept e which is related to both c and d.

The rationale for this new relation is the following: usually, in the definition of a city we can find the country to which it belongs; or in the definition of a country we can find the neighbouring countries. Therefore, if two locations are related (e.g. if one is inside the other, or if they are near each other) then it is very likely that one will appear in the definition of the other. Furthermore, in a web page about a location, usually some nearby location is cited, so we can infer whether the web page is really referring to the same location that we want.

For example, the following is the definition of *Chicago* in WordNet:

1. Chicago, Windy City – (largest city in Illinois; located on Lake Michigan)

Therefore, there are gloss relationships between *Chicago*, *Illinois* and *Lake Michigan*.

From the original documents provided by the user, the system collects all the entities located in them, and their frequencies of appearance. These concepts can now be arranged in several connected subgraphs corresponding to the equivalence classes produced by the relationship $\mathcal{G}$.

The following is the connected subgraphs that were obtained for the concept *Valparaiso*, looking at other locations that appeared near it in *The Voyages of the Beagle*:

{(Mexico,1)},
{(Saint Lucia,1)},
{(Guayaquil,2)},
{(Australia,1)},
{(Talcahuano,3)},
{(Lisboa,1)},
{(Bahia,1)},
{(Buenos Ayres,1)},
{(Mendoza,1)},
{(Copiapo,7)},
{(Calabria,1)},
{(S. Carlos,1)},
{(Chimborazo,1)},
{(Coquimbo,5)},
{(Banda,1)},
{(Chonos Archipelago,1)},
{(Aconcagua,3)},
{(England,1)},
{(Madeira,1)},
{(Callao,4)},
{(Chiloe,2),(Brasil,1),(Central America,1),(Tierra del Fuego,1),(Lima,3),(Peru,1),
(America,1),(Patagonia,3),(Valparaiso,29)(Concepcion,5),(Gran Santiago,3),
(South America,3),(Chile,10),(Pacific,1)}

We define the weight of a subgraph as the sum of all the frequencies of the entities in it. As can be observed, the last subgraph contains many more concepts than the rest, and with the highest weights. To perform the filtering, for each document downloaded from the Internet the same process is performed: all the entities found in those documents are selected, and the subgraphs are generated in the same way. The filtering heuristic we used consists in comparing the most weighty subgraphs: the one from the original documents and the one from the web page. The page is kept only if the most weighty subgraphs have shared locations. Otherwise, it will be filtered out and its contents will not be considered for the summary.

For illustration, the following are the connected subgraphs for a document that refers to *Valparaiso* in the United States, instead than in Chile. The connected subgraph with the highest weight here is the one with refers to the United States (with total weight 15); while the subgraph referring to Chile only has weight 1. The document was rejected.

{(Columbia,1)}, {(Brasil,1)}, {(Chile,1)},
{(Cornhusker State,1)}, {(Chicago,2),(Lake Michigan,2)},
{(Middle West,1),(Detroit,1),(Everglade State,1),(Mexico,1),(South Bend,1),
(Hoosier State,3),(United States,7)}

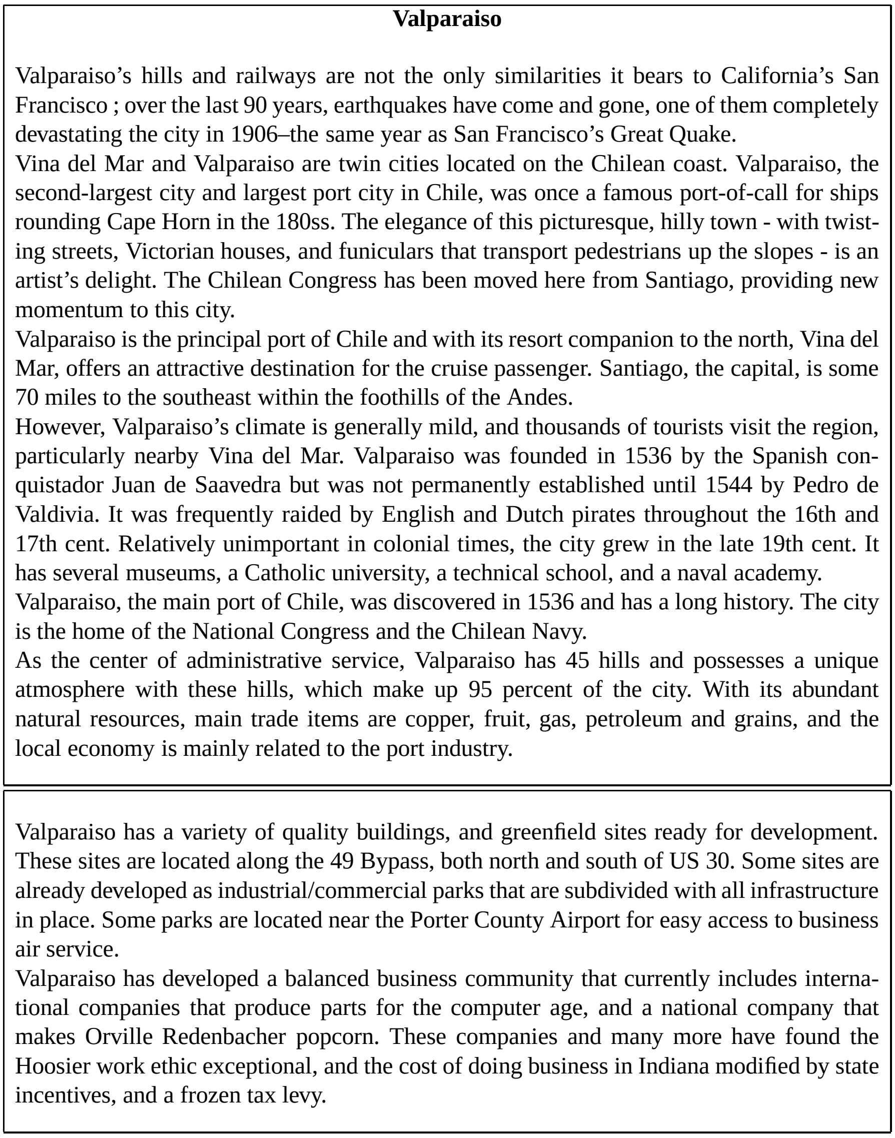
2.4 Paragraph filter and result generation

The first filtering helped eliminate much irrelevant information, but still many of the web pages contained information that was not useful for generating a summary page. Some of the pages contained special offers from travel agencies; other were travel journals written by people, offering their subjective opinions, but without much interest; and others referred to organisations that were related to the concept, such as *the University of Valparaiso*.

The following two heuristics were now applied in order to retain only the most informative paragraphs:

- Firstly, it was required that the term had to appear as head of a Noun Phrase, not as a modifier. For example, in *Valparaiso Authorities*, *Valparaiso* is modifying the head of the Noun Phrase *Authorities*, and hence it was not considered relevant.
- The second heuristic looks for sentences where the term appears in the subject position, and then keeps all the subsequent sentences up to the next end-of-paragraph. If the concept does not appear in the subject position, then the paragraph is ignored. This heuristic allowed us to rule out more than 75% of the pages that did not contain relevant information, such as travel logs to a particular location, or travel agency advertisements.

Figure 2 shows the summary page for the city *Valparaiso*, generated after all the processing.



Valparaiso

Valparaiso’s hills and railways are not the only similarities it bears to California’s San Francisco ; over the last 90 years, earthquakes have come and gone, one of them completely devastating the city in 1906–the same year as San Francisco’s Great Quake.
Vina del Mar and Valparaiso are twin cities located on the Chilean coast. Valparaiso, the second-largest city and largest port city in Chile, was once a famous port-of-call for ships rounding Cape Horn in the 180ss. The elegance of this picturesque, hilly town - with twisting streets, Victorian houses, and funiculars that transport pedestrians up the slopes - is an artist’s delight. The Chilean Congress has been moved here from Santiago, providing new momentum to this city.
Valparaiso is the principal port of Chile and with its resort companion to the north, Vina del Mar, offers an attractive destination for the cruise passenger. Santiago, the capital, is some 70 miles to the southeast within the foothills of the Andes.
However, Valparaiso’s climate is generally mild, and thousands of tourists visit the region, particularly nearby Vina del Mar. Valparaiso was founded in 1536 by the Spanish conquistador Juan de Saavedra but was not permanently established until 1544 by Pedro de Valdivia. It was frequently raided by English and Dutch pirates throughout the 16th and 17th cent. Relatively unimportant in colonial times, the city grew in the late 19th cent. It has several museums, a Catholic university, a technical school, and a naval academy.
Valparaiso, the main port of Chile, was discovered in 1536 and has a long history. The city is the home of the National Congress and the Chilean Navy.
As the center of administrative service, Valparaiso has 45 hills and possesses a unique atmosphere with these hills, which make up 95 percent of the city. With its abundant natural resources, main trade items are copper, fruit, gas, petroleum and grains, and the local economy is mainly related to the port industry.

Valparaiso has a variety of quality buildings, and greenfield sites ready for development. These sites are located along the 49 Bypass, both north and south of US 30. Some sites are already developed as industrial/commercial parks that are subdivided with all infrastructure in place. Some parks are located near the Porter County Airport for easy access to business air service.
Valparaiso has developed a balanced business community that currently includes international companies that produce parts for the computer age, and a national company that makes Orville Redenbacher popcorn. These companies and many more have found the Hoosier work ethic exceptional, and the cost of doing business in Indiana modified by state incentives, and a frozen tax levy.

Fig. 2. Generated page about the concept *Valparaiso*. The two paragraphs below were discarded by the page filter because they refer a different city in the U.S.

Concept	Valparaiso	Cordillera	Patagonia	Buenos Ayres	Chiloe
Pages downloaded	100	89	95	91	98
Pages with information	18	8	11	2	13
Pages selected	13(0)	2(0)	5(0)	2(2)	8(0)
Relevant Paragraphs	29	14	11	2	34
Paragraphs Selected	16(0)	3(0)	5(0)	2(2)	13(0)
Page recall	72.22%	25%	45.45%	0%	61.54%
Page precision	100%	100%	100%	0%	100%
Paragraph recall	55.17%	21.43%	45.45%	0%	38.24%
Paragraph precision	100%	100%	100%	0%	100%

Table 1. Results for the multi-document summariser for information collected from the Internet. Between parenthesis are the number of selected pages and paragraphs that were incorrect.

Error type	**Number of times**
Concept not as subject	12
Syntax analysis error	11
Unknown synonym in the text	8
Incorrect processing of other languages	4
Indirect information	2
Pronoun coreference not solved	1
The whole page was filtered out	1

Table 2. Reasons for some of the recall errors committed by the algorithm.

3 Evaluation

Retrieval algorithms are usually evaluated in terms of *recall*, the percentage of relevant information that was found, and *precision*, the percentage of the selected information that was really relevant. Our approach, which consists of successive filterings of the web pages and their paragraphs, is intended to maximise precision, even though recall may suffer a little if relevant pages are filtered out.

Table 1 shows the results for five different relevant concepts extracted from *The Voyages of the Beagle*. Nearly 500 documents retrieved from the Internet were examined by hand, and all their paragraphs were annotated as containing or not relevant information about those terms. There were, in mean, 67 paragraphs per web page; table cells and list items were considered separate paragraphs. The first five lines describe:

- The number of pages that were downloaded for each concept.
- The number of pages that contained relevant information.
- Pages selected; between parenthesis, how many of them are errors.
- The number of paragraphs that contained relevant information.
- The number of paragraphs selected; between parenthesis is the number of mistakes.

With this information, it is possible to calculate the recall and precision concerning the choice of web pages and of paragraphs in them. These values appear in the last four lines in the table.

Table 2 describes some of the errors that provoked the drop in recall (the paragraphs that were missed). Most of the errors were due to the following three causes: the concept not appearing as subject in a relevant paragraph; parser errors; and the use of unknown synonyms of the relevant term. For example, in several occasions *Cordillera* (the Spanish for *mountain ridge*, with which Darwin refers to the Andes mountains) was referred to as *Andes* in the web pages, and the system did not recognise it. A few number of mistakes happened because of various other reasons.

3.1 Discussion

Summaries were automatically collected for the 131 terms that were identified automatically in the text. These represent locations, such as as *Spain*, *Scotland* or *Germany*; people, such as the *Captain FitzRoy* or *York Minster* (a Fuegian that travelled in the ship); bodies of water (rivers, lakes, seas); and some artifacts (e.g. *lazo*) and animals (e.g. *Carrancha* and *bizcacha*). Of the 131 summaries generated in total, 37 were left empty, because no relevant information was found for them in the Internet pages. Either no web page was found, all of them were filtered, or the term never appeared in subject position.

The mean time to collect a summary was 20 minutes, of which 30 seconds were necessary to study the term (study its context in the original documents and create the connected subgraphs using *gloss* links); eighteen minutes in average were needed to download the one hundred documents from the Internet and process them with the linguistic tools; and one minute and a half was needed in order to analyse the texts and produce the summary file.

The contextual clues taken from the source document were really useful for filtering the downloaded pages. This was specially important about concepts that can have multiple meanings. For example, the term *Cordillera* is the Spanish for *mountain ridge*, and there were pages about many ridges in the world, but all the paragraphs retained correctly refer to the Andean mountains. The filtering that is done with the sub-graphs also ruled out a Catalan wine called *Cordillera*, and a church and a Chilean province with the same name.

The heuristics were in general good for avoiding pages about travel accounts, travel agency offers, hotels, and other spurious information. There was one weak point, though. When an unknown term is really very polysemous, as it is the case of some locations that refer to many different places, then the filtering performed is not accurate enough. That was the case of *St. Helena* or *La Plata*, for which there exist many places with the same names, and much of the data collected was not about the places Darwin visited.

A form was provided to twelve users of the information system, in order to discover their opinion about this functionality. They were asked to rate the collection of additional information in a scale from 1 to 5. The average answer was 4.67, with a standard deviation of 1.83, which shows that there was general agreement about its usefulness.

4 Conclusions and Future work

A new algorithm has been implemented that collects information about the relevant terms found in a document, or specified by a user. Several filters ensure as much as possible that the quality of the obtained information is good. The filters that have been used helped eliminate most of the irrelevant data. When applied to Darwin's *The Voyages of the Beagle*, it automatically generated summary pages for 131 terms, which include locations and people names, artifacts, animals and bodies of water. 37 of them were left empty, as no relevant information was found for them, and the remaining summaries have a length ranging from a few sentences, up to nearly 18,000 words, in the cases of *Chile* and *Brazil*.

A manual evaluation of the results shows that precision is very high in most of the summaries, reaching 100% in many of them. Recall can still be improved, as many relevant paragraphs were rejected by the filters; it ranged from 20% to 60% in most of the cases evaluated. Most of the errors were due to syntax analysis errors, and to several occasions in which the heuristic of looking for the term at the subject position failed.

The method for generating summaries can be further improved. The following are some ideas that may be addressed in the future:

- A reordering of the paragraphs inside the final document according to their topic, so those which are similar appear together.
- A merging of the paragraphs that convey the same information. Sometimes there are two paragraphs that are nearly identical, except for one or two words, or for the punctuation symbols. In these cases, the repeated information should be discarded.
- A better way to identify synonyms of the terms that may appear in the pages, and some co-reference resolution for pronouns.

References

E. Alfonseca and S. Manandhar. Extending a lexical ontology by a combination of distributional semantics signatures. In *Lecture Notes in Artificial Intelligence 2473. Springer Verlag*, 2002.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

D. Boley, M. Gini, R. Gross, E.-H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Document categorization and query generation on the world wide web using WebACE. *AI Review*, 13(5–6):365–391, 1999.

T. Brants. *TnT - A Statistical Part-of-Speech Tagger*. User manual, 2000.

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the WWW. *Artificial Intelligence*, 118(1–2), 1999.

M. Diligenti, F. Coetzee, S. Lawrence, C. Lee Giles, and M. Gori. Focused crawling using context graphs. In *26th Int. Conf. on Very Large Databases, VLDB 2000*, pages 527–534, 2000.

R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. University of sheffield: Description of the lasie system as used for muc-6. In *MUC-6*, pages 207–220, 1995.

Steve Lawrence. Context in web search. *IEEE Data Engineering Bulletin*, 23(3):25–32, 2000.

A. McCallum, K. Nigam, J. Rennie, and K. Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI-99*, 1999.

G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *Third ACL Workshop on Very Large Corpora*, pages 82–94. Kluwer, 1995.

J. Rennie and A. K. McCallum. Using reinforcement learning to spider the Web efficiently. In I. Bratko and S. Dzeroski, editors, *Proceedings of ICML-99*, pages 335–343, 1999.

E. M. Voorhees and D. K. Harman. *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)*. Department of Commerce, NIST, 2001.