The Gaps of the Thesaurus Wordnet Used in Information Retrieval

Javier de la Mata, Jose A. Olivas, and Jesús Serrano-Guerrero

University of Castilla-La Mancha
Paseo de la universidad 4,
13071, Ciudad Real, Spain
{JavierL.Mata, Jesus.Serrano3}@alu.uclm.es,
Joseangel.olivas@uclm.es

Abstract. Due to the exponential growth of Internet it is very important to have good knowledge structures that let to obtain good results in Web search. The aim of this work is to discover the user tendencies when they use the search engines and to know the limitations of the knowledge structures that GUMSe[1] uses. With this information is possible to design a more efficient system. For this reason, it is analyzed the set of keywords and queries more frequently used in the search engines and how WordNet manage it. This information is very useful to avoid bad situations in our meta-search engine.

1 Introduction

Different users may differ about the relevance of several documents obtained using the same query. Relevance is a subjective notion. Standard search engines try to solve the main problems that affect the quality of the results with the aim of obtaining a relevant collection of documents. The main sources of these problems are the ambiguity and the vocabulary. But the search engine needs to know some kind of semantic information that let it to improve the results.

The main sources that are usually used to discover the semantic information and relations are the dictionaries, thesaurus or ontologies. WordNet [1] is one of the main tools used in information retrieval processes, mainly for disambiguation tasks. These tools have several problems such as for example the granularity of the senses [2] or the lack of recent terms. For example, WordNet 2.0 doesn't recognize the terms "XML" or "CORBA". Other times, it recognizes the common senses of one word, like in the case of "SOAP", but does not recognize the new sense (Simple Object Access Protocol for the previous case).

This work is focused on the study of the weak points of WordNet. Our objective is to use its information to improve our knowledge structures. For this reason, the study of the user's query is an important aspect that helps us to know the terms not recognized by WordNet. This information can be later on used to get better results.

There are many works focused on the study of user queries. Most of them offer statistics about the number of keywords, the number of queries per session, and other

[1] GUMSe: Gum Search, meta search engine Developer in the framework of SMILe-ORETO-UCLM (Soft Management of Internet e-Laboratory) research group.

R. Moreno Díaz et al. (Eds.): EUROCAST 2005, LNCS 3643, pp. 229 – 234, 2005.

statistical measures [3]. For example, Jansen & Spink [4] studied the queries of the users for the Excite Search Engine. But this study was concentrate on users' sessions, queries and terms. Other interesting study is the comparison of three different search mechanisms: query-based search, directory-based search and phrase-based query reformulation assisted search [5]. This study concludes that query reformulation can significantly improve the relevance of the documents but with an increase in the search time and the cognitive load.

Usually one keyword is considered a unique word. It is a serious problem because in many situations it is not possible to use an isolated word to describe a text. For this reason, in this work, it is defined a 'keyword' as a word or combination of words that describe a remarkable characteristic or item of one topic. But, usually user queries have several words, and now another problem appears: how to distinguish keywords within a query?

GUMSe [6] have been developed like a platform that allows us to test new ideas in Web search processes. Using the classic technique of query expansion, GUMSe semantically obtain a collection of additional queries related to the original one. New queries are generated replacing or introducing new related terms to previous ones by means of synonymy, hyponymy or hyperonymy relations.

2 Methodology

The first step was to obtain a collection of user queries. Nowadays, the system is in test phase and we do not have enough queries to make an exhaustive study. Nevertheless, there are many web sites that make available the most popular queries that users submit, or even all the queries. For the object of this study, the main source of our collection of keywords was Hitbrain[2] and MetaSpy. The Hitbrain Web site offers a collection of 10.000 keywords and information about each one such as the frequency of use, the position in the monthly ranking and the last positions. This site assumes that a keyword can consist of several words. In addition, we used other sets of queries from MetaSpy. At this point we have to distinguish between keywords and queries. The difference between both is a little bit unclear because one keyword is also a query, but a query is not a keyword. That is to say, a query can be formed by one or more keywords.

The following step was to adapt the data from different sources to the same format for its later processing and study. Once it was completed, we made three different experiments:

1. **Study of the terms recognized by WordNet**: counting the number of terms that WordNet recognizes.
2. **Study of the terms recognized by WordNet with bad sense**: it is analyzeed if WordNet recognizes the terms in a wrong way. Frequently WordNet recognizes one keyword in a wrong way due to the polisemy of the terms.
3. **Study of the topics of the terms**: Finally, it is analyzed what topics are most frequently used by users.

[2] http:/www.hitbrain.com

3 Results

3.1 Terms Recognized by WordNet

Our first experiment was to count the number of keywords that WordNet recognizes. In this experiment, three different situations appear: WordNet recognizes the keyword, WordNet does not recognize the keyword and WordNet recognizes the keyword in a wrong way. In this analysis, the third case is not considered. We assume that WordNet recognized the items in a right way.

For this study two test collections of terms were used: Col1 and Col2. The first collection (Col1) from Hitbrain was compound by 10,007 keywords. The second (Col2) was a collection of queries obtained from Metaspy with 123,809 queries.

Table 1. Characteristics of the two collections

	Items	*Type*	*Source*
Col1	10,007	Keywords	HitBrain
Col2	123,809	Queries	MetaSpy
Col2	307,286	Keywords	MetaSpy

The results of the analysis of Col1 are shown in table 2. This collection has keywords formed by one to five words. The experiment reveals that WordNet recognizes the 66% of the keywords formed by one word, and this percentage decrease when the number of words that compound each keyword increase. WordNet only recognizes the 4,53% of the keywords formed by 2 words (see Table 2), and practically the 0% of the keyword with more than 2 words. Total: the 45,45% of the keywords are recognized by WordNet, where the 66,6% of them are keywords with only one word.

Table 2. Results of the analysis for Col1. The table shows the number of words that form one keyword, the number of keywords recognized and not recognized by WordNet, and the percentage of keywords recognized.

Nº WORDS	*RECOGNIZED*	*NOT RECOGNIZED*	*Nº KEYWORDS*	*PERCENTAGE*
1 word	4420	2250	6670	66,26%
2 words	127	2676	2803	4,53%%
3 words	0	451	451	0%
4 words	1	61	62	1,61%
5 words	0	13	13	0%
Others	0	8	8	0%
TOTAL	4548	5459	10007	45,45%

The analysis of Col2 was different because it is formed by queries of the users, and not only keywords. One query can be formed by one or more keywords. In this experiment we use two criteria of evaluation of the query: the first one was to consider the query like a keyword (formed by one or more words) and the second was to consider one query formed by one or more keywords (each word is assumed like a keyword).

Table 3. Results of the analysis of Col2

Case	*Items*	*Rec.*	*Not Rec.*	*Per.*
Queries	123,809	31,637	92,172	25,55%
Keywords	307,286	152,679	154,607	49,69%

The results in both cases are very different. For the first case the number of queries was 123,809. The number of queries recognised was very low (only the 25,5%) and it was only the queries formed by one keyword. In the second case each word is considered like a keyword. The number of words processed was 307,286 (the average number of terms by query was 2,48) and WordNet recognized the 49,68% of the words. This result is very similar to the average number of the first experiment.

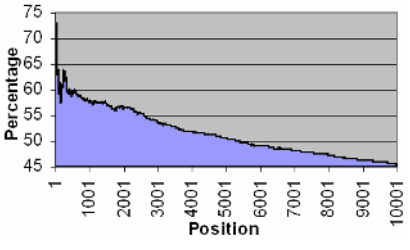

Fig. 1. Average of the terms recognized by WordNet for Col1 related to the frequency of use

Another important aspect is that WordNet recognized around the 60% for the 600 first keywords (see Figure 1). But the average is decreasing in the keywords with lower position (or frequency of use). This means that if the number of different keywords is very small, then the behaviour of WordNet is better than if there are many keywords. But if the frequency of the use of the first keywords is very high, then the behaviour of WordNet improves because correct cases are more frequents. It implies that it could be better to improve only the behaviour for the keywords more frequently used because they are the more probable situations.

3.2 Terms Recognized by WordNet with Bad Sense

Next work is a preliminary study of the precision of WordNet. This experiment uses the 250 first keywords of the collection Col1, where WordNet recognized only 180. The process consists of verifying what keywords recognized by WordNet are wrong. For the accomplishment of this study, the meanings of each keyword recognized by WordNet were observed. If WordNet has the correct sense then the keyword has been recognized right. On the contrary, the sense has been selected incorrectly. Evidently this test is subjective, since the criterion to decide if one keyword is wrong depends of the person that makes the test. For this reason this aspect of our investigation can be improved in the future. The results of this experiment are showed in table 4.

Table 4. Results of the second experiment

Nº Keywords	*180*	
OK	157	87,2%
WRONG	23	12,8%

This experiment shows that around the 12% of the keywords recognized by WordNet are not in the correct sense. This is the case, for example, of the keyword "amazon" that in WordNet can be: "a large strong and aggressive woman", "one of a nation of women warriors of Scythia", "a major South American river" or "mainly green tropical American parrot". The previous meanings are correct and there are people that looking for these topics, but in Internet, the usual case (we think) is to use this keyword to search a web site that sells books.

3.3 Study of the Topics of the Terms

The last experiment classifies the keywords with the objective of knowing what domains are more demanded for the users. This information is useful to know why the WordNet thesaurus fails. In this experiment we analyze the first 250 keywords of Col2 and each keyword was assigned to one or more pre-established categories. Table 5 shows the 10 categories. The "other" category includes the keywords that are not in the previous nine ones. This experiment uses, such as the previous one, a subjective criterion. The results show that many queries in Internet are about Internet. This causes that the queries about Internet are not recognized in some cases in a correct way by WordNet. An exhaustive analysis can help us to know the weakness of WordNet and what aspects are necessary to improve if we want to get a more efficient meta-search engine.

Table 5. Distribution of the keywords by categories and wrong senses in each one

Category	*Hits*	*Errors*	*% OK*	*% ERROR*
Web	75	10	30%	13,3%
Computer	55	7	22%	12,7%
Location	25	4	10%	16%
Games	21	3	8,4%	14,3%
Music	19	2	7,6%	10,5%
Movies	15	1	6%	6,6%
People	13	0	5,2%	0%
Sport	10	0	4%	0%
Health	6	0	2,4%	0%
Others	99	0	39,6%	0%

It is also studied the relation between the categories and the recognition of bad senses in WordNet. This experiment reveals that the two main categories that causes fails are *Web* and *Computer* categories (the category *Others* is not considered).There are also other categories such as *Music* or *Games* that are in continuous change. This situation make difficult to update a knowledge structure and it is so easy that it fails when terms about these categories are used.

4 Conclusions

In this study, WordNet recognizes the 45,45% of more frequently keywords used in user queries. The study used 10.007 keywords. The keywords have 1 to 5 words. Other important aspect is that there are situations where WordNet recognizes keywords, but in a wrong way. These cases are very infrequent but they can produce mistakes in search processes.

It is necessary to have good knowledge structures that could be used for information retrieval purposes to focus the search and to obtain better results. For this reason, a subsystem specialized in the improvement of the knowledge structures could be a good support for search engines.

References

1. Barto, A. G., Sutton, R. S., Anderson, C. W.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Systems, Man, and Cybernetics, Vol. SMC-13, (1983) 834–846.
2. Agirre, E, Lopez de Lacalle, O.: Clustering Wordnet word senses. In: Proceedings Recent Advances on Natural Language Processing (RANLP'03) (2003)
3. Werbos, P. J.: Neural networks & the human mind: New mathematics fits humanistic insight. In: Proceedings of the 1992 IEEE Trans. Systems, Man, and Cybernetics, Vol. 1, (1992) 78–83.
4. Jansen, B. J., Spink, A., Saracevic, T.: Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing & Management, Vol. 36, No. 2, (2000) 207-227.
5. Bruza, P., McArthur, R., Dennis, S.: Interactive internet search: Keyword, directory and query reformulation mechanisms compared. In: Proc. 23rd annual Int. ACM SIGIR conf. on Res. and Devel. in Inform. Retrieval. Athens, ACM Press: New York. (2000) 280-287.
6. Olivas, J. A., de la Mata, J., Serrano-Guerrero, J.: Ontology Constructor Agent for improving Web Search with GUMSe. In: Proc. of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'04), Vol. 2, Perugia, Italy, (2004). 1341-1348.