Hyponymy Extraction and Web Search Behavior Analysis Based On Query Reformulation

Rui P. Costa and Nuno Seco

Cognitive and Media Systems Group, CISUC
University of Coimbra, Portugal
racosta@student.dei.uc.pt, nseco@dei.uc.pt

Abstract. A web search engine log is a very rich source of semantic knowledge. In this paper we focus on the extraction of hyponymy relations from individual user sessions by examining, search behavior. The results obtained allow us to identify specific reformulation models as ones that more frequently represent hyponymy relations. The extracted relations reflect the knowledge that the user is employing while searching the web. Simultaneously, this study leads to a better understanding of web user search behavior.

Key words: Semantic Web Usage Mining, Query Reformulation, Hyponymy Extraction, Web User Search Behavior.

1 Introduction

Web Usage Mining applies concepts from the field of data mining to the information banked on the Web (e.g. to the logs of a web search engine). In this study we consider a log to be "an electronic record of interactions that have occurred during a searching episode between a web search engine and users searching for information on that web search engine" [1]. Web Usage Mining is a new field fueled by a huge amount of data, an amount that grows every second, for example, the data that Google stores in its logs every second. Taking into account the fact that it is mainly humans that conduct the search requests contained in these logs, this work attempts to extract a representation of the knowledge being used by looking into search behaviors. Log analysis fits well in this type of research because it provides the "most naturally-occurring and large-scale data set of query modifications" [2].

The driving assumption in this paper is that there are semantic relations between the terms contained in a session. The semantic relation studied is the hyponymy.

Many statistical studies have been conducted (as can be found in [1,3]), but very few have used logs from a generic web search engine. In our study the logs are from a generic search engine, thus providing an interesting longitudinal study [3].

A previous study has looked at ways of building taxonomies given an initial set of 14 main categories [4]. For every query issued they analyzed the category of pages returned and added the query terms to that category. Another study related to user behavior and semantics concluded that by using a specific extraction methodology an ontology may be extracted [5].

Several studies have been concerned with hyponymy extraction [6,7] and ontology learning [8]. These studies usually apply lexico-semantic patterns (e.g. *is a* and *such as*) in order to extract hyponymy relations and build ontologies.

Regarding query reformulation, one study examined multiple query reformulations on the Web in the context of interactive information retrieval [2]. The results indicate that query reformulation is the product of user interaction with the information retrieval system. They used a log that corresponded to a particular day, cleaned it manually and obtained 313 search sessions. However, results regarding semantics were not presented. This study is further explained in section 2.

Two studies focused on the Portuguese language and web search logs. In the work of [9], 440 people were emailed from a Computer Science Department in a Brazilian University, and asked to keep track of their queries during a one month period. The purpose was to study the application of natural language in formulating queries. The second work studied the identification of user sessions in a generic log with the intent of conducting future work enabling the extraction of semantic relations [10]. Regarding the Portuguese language we did not found any study dealing with the semantic aspects.

This paper proposes a new approach for hyponymy extraction and gives some clues towards the understanding of typical web user search behaviors, both based on query reformulation within individual user sessions. The relations extracted can enrich well-known ontologies like Wordnet[1] with new hyponymy relations. Considering that knowledge is growing every day this may provide an interesting way of updating ontologies, as we believe that this new knowledge will be reflected in the logs.

This paper is organized in the following manner: Section 2 describes a theoretical framework for web user search behavior, in Section 3 the hyponymy extraction method is explained, in Section 4 preliminary experiments are presented, and finally in Section 5 we conclude the paper.

2 Web User Search Behavior

The theoretical framework of interactive information retrieval that was applied was the one from Rieh [2] that derives from Saracevic [13]. Saracevic defines interaction as the "sequence of processes occurring in several connected levels of strata" ([13] p. 316). As can be deducted from the Figure 1, the interaction between the user and the system is complex. On the users side, there are three levels: cognitive (interaction with texts and their representations), affective (interaction with intentions) and situational (interaction with given or problems-at-hand). On the other side, we have the computer, where there are the engineering, processing and content levels. Both sides meet at the surface level where adaptation through feedback occurs.

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

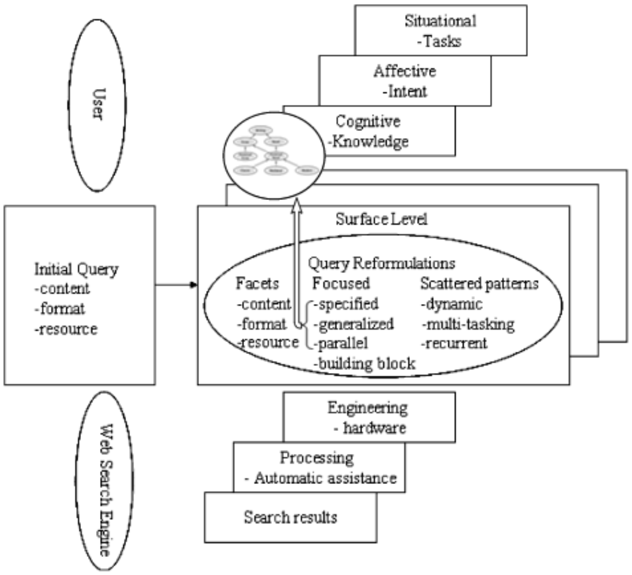

Fig. 1. Model of web query reformulation [2] (with user knowledge extraction link).

A unique, non-intrusive way to understand the user (cognitive, affective and situational facets) is by web log mining, but the current search engines do not seem to use this to their advantage. Rieh [2] reveals three main categories of reformulation within user sessions:

- *content* - 80.3% (changes to the meaning of a query);
- *format* - 14.4% (changes without altering the meaning of the query);
- *resource* - 2.8% (changes in types of information resources (e.g. news and images)).

The category studied in this paper is content since, it is the most frequent category and the most related to our study. The sub-facets studied are: specialization, generalization and parallelization. Specialization corresponds to the addition model, generalization to the deletion model and finally parallelization to the substitution model (see section 3). Figure 1 was updated from [2] with a link between query reformulation and user knowledge.

Our conclusions are in contrast to [14,15] who comment that transaction logs can only deal with the actions of the user, not their perceptions, emotions or background. Our study proposes a new method that can be used to extract a portion of the user knowledge (hyponymy relations).

Table 1. Three queries that exemplify a session

IP	Search URL	Date/Time
127.0.0.1	GET /pesquisa?lang=pt&index=sidra&terms=**virtual**&**books**	01/Oct/2003:00:00:07
127.0.0.1	GET /pesquisa?lang=pt&index=sidra&terms=**free**&**virtual**&**books**	01/Oct/2003:00:01:07
127.0.0.1	GET /pesquisa?lang=pt&index=sidra&terms=**virtual**&**libraries**	01/Oct/2003:00:10:47

Table 2. User reformulation model of addition (specialization)

Term(s) addition **beginning/middle/end**	This behavior happens when the user wants to specify the search.
Patterns	1. The new expression is a hyponym of the old one. 2. The old expression is a hyponym of the terms entered. 3. The new expression is a hyponym of the terms entered. 4. The new expression is a hyponym of the terms before the new term (just for middle case). 5. The old expression is a hyponym of the terms before the new term (just for middle case).
Examples	1. seguro (insurance), seguro escolar (scholar insurance). *Conclusion*: Insurance scholar is a hyponym of insurance. 2. pinheiro (pine), pinheiro árvore (pine tree). *Conclusion*: Pine is a hyponym of tree. 3. nações unidas (united nations), organização das nações unidas (united nations organization). *Conclusion*: United nations organization is a hyponym of organization. 4. dia sem carros (day without cars), dia europeu sem carros (european day without cars). *Conclusion*: European day without cars is a hyponym of day. 5. região oeste (west region), região turismo oeste (west region tourism). *Conclusion*: West region tourism is a hyponym of region.

3 Hyponymy Extraction using User Reformulation Models

In the present study we consider a session as a sequence of queries that the user issues to satisfy his information needs [10]. An example of a session is shown in Table 1.

The algorithm used to detect sessions is presented in [10]. Our session detection algorithm used a 15 minute time limit as the maximum duration of a

Table 3. User reformulation model of deletion (generalization)

Term(s) deletion **beginning/middle/end**	This behavior happens when the user wants to generalize the search.
Patterns	1. The old expression is a hyponym of the new one. 2. The new expression is a hyponym of the deletion term. 3. The old expression is a hyponym of the deletion term. 4. The old expression is a hyponym of the expression before the deletion terms (just for middle case). 5. The new expression is a hyponym of the expression before the deletion terms (just for middle case).
Examples	1. planeta terra (planet earth), planeta (planet). *Conclusion*: Planet earth is a hyponym of planet. 2. orientação desporto (orientation sport), orientação (orientation). *Conclusion*: Orientation is a hyponym of sport. 3. parque de campismo (camping park), campismo (camping). *Conclusion*: Camping park is a hyponym of park. 4. tribunal penal internacional (international criminal court), tribunal internacional (international court). *Conclusion*: International criminal court is a hyponym of court. 5. desenvolvimento ambiental sustentável (sustainable ambiental development), desenvolvimento sustentável (sustainable development). *Conclusion*: Sustainable development is a hyponym of development.

session. This threshold is based on the work of [11] where it was concluded that the optimal session interval is somewhere within the range of 10 to 15 minutes.

Hyponymy is the semantic relation of being subordinate or belonging to a lower rank or class; for instance apple is a fruit.

The new method that we propose for hyponymy extraction must be applied on a per session basis and on each two sequential queries. The user reformulation models proposed are based on the work of [12], where three different patterns of query reformulation were identified: substitution, addition and deletion.

Tables 2, 3, 4 and 5 explain the three user reformulation models studied (addition, deletion and substitution). Different positions of query reformulation (at the beginning, in the middle and at the end) were studied. Each table contains several hyponymy extraction patterns and the respective examples. These patterns were identified during manual log analysis. The examples given are extracted from the logs of the Tumba[1], a search engine focused on Portuguese

[1] http://www.tumba.pt

Table 4. User reformulation model of substitution (parallelization)

Term(s) substitution beginning/middle/end	This behavior happens when the user wants to do a parallel move (co-hyponymy).
Patterns	1. The new expression is a hyponym of the constant terms. 2. The old expression is a hyponym of the constant terms.
Examples	1. agência de actores (actors agency), agência de publicidade (advertising agency). *Conclusion*: Advertising agency is a hyponym of agency. 2. agência de actores (actors agency), agência de publicidade (advertising agency). *Conclusion*: Actors agency is a hyponym of agency.

Table 5. User reformulation model of total substitution (generalization | specialization)

Total substitution	This behavior happens when the user wants to specify or generalize.
Patterns	1. The old expression is a hyponym of the new one (generalization). 2. The new expression is a hyponym of the old one (specification).
Examples	1. mba (mba), pós graduação (pos graduation). *Conclusion*: Mba is a hyponym of pos graduation. 2. anti virus (anti virus), norton (norton). *Conclusion*: Norton is a hyponym of anti virus.

language. Therefore the reformulation patterns are only applicable to Portuguese (the English translation is given to assist the reader).

4 Experiments

The data used corresponds to 1,7 million cleaned query entries from 2003 from the Tumba. The log was cleaned by removing all entries that did not correspond to true search queries (e.g. bots and watchdog server queries). We found 75320 distinct sessions after applying the method presented in Section 3.

The following two topics present different evaluation strategies. The first does not need human intervention while the second is completely manual.

Using an approach similar to [16] we used Google in order to evaluate the extracted hyponymy relations. The idea is simple, using the extraction method presented and a pattern between the terms extracted it is possible to validate semantic relations. The pattern is linked by the semantic relation. For instance, if we have the terms *mouse* and *animal*, and the pattern ”is an” (for hyponymy) between them, Google returns 900 results. Therefore, we can say that *mouse* is an hyponym of *animal* with a frequency of 900. However, there are sometimes false positives, mainly because of the missing words. For instance, *education is a ministry* has a frequency of 7240, but the correct form would be *ministry of education is a ministry*. This problem can be smoothed by adding articles before the terms (e.g. *an education is a ministry*, returns 0 results).

To reinforce the results evaluated from the Google search, a sub-set of 30 hyponym relations were randomly chosen. For each sub-set we selected a number of pattern examples proportional to the total amount of each pattern. When the patterns did not have at least 30 relations, we used logs from others years. However, for the substitution at the beginning, the maximum number of possible relations was 25. The evaluation was done by 10 people. For each relation the evaluator had to choose one of the following options: wrong, some sense or correct. The “some sense” option allows the evaluator to state that s/he is not sure about the relation, making the study more realistic.

Table 6. General model results for hyponymy extraction

Model	**Model amount**	**Google**	**Correct**	**Some sense**	**Wrong**
Addition end	17073	231	55,33%	21,67%	23%
Addition beginning	5665	163	58%	20,67%	21,33%
Addition middle	3421	52	89,33%	7,33%	3,33%
Deletion end	7281	107	53,66%	19,67%	26,67%
Deletion beginning	7170	49	64,67%	25,34%	10%
Deletion middle	3333	49	89,66%	8,34%	2%
Substitution end	27755	553	88%	12%	0%
Substitution beginning	15767	21	32,5%	21,66%	45,83%
Substitution middle	2286	27	64%	14%	22%
Total substitution	70382	41	38,34%	30,67%	31%
Total	160133	1295	61,92%	19,58%	18,5%

We used Google to evaluate our method, and manual verification to evaluate Google in order to ensure a threshold of confidence regarding the results given by Google. Table 6 shows the results from the experiments using the Tumba logs from 2003. The third column holds the number of unique relations that Google evaluated and the subsequent columns hold the manual evaluation results. Table 7 shows the quantity and evaluation of each pattern within each model. For

Table 7. Model patterns results. Each cell contains the total number of hyponymy relations validated by Google search and within parenthesis the manual validation results (both for a specific pattern).

Model	**1st** (C;SS;W)	**2nd** (C;SS;W)	**3rd** (C;SS;W)	**4th** (C;SS;W)	**5th** (C;SS;W)
Addition end	76 (82;15;3)	152 (35;23;42)	3 (53;40;7)		
Addition beginning	0	138 (38;30;32)	25 (99;1;0)		
Addition middle	0	2 (45;15;40)	1 (50;40;10)	30 (96;4;0)	19 (92;7;1)
Deletion end	48 (84;13;3)	48 (32;24;44)	1 (70;30;0)		
Deletion beginning	0	28 (53;35;12)	21 (78;14;8)		
Deletion middle	0	0	0	39 (90;8;2)	10 (90;10;0)
Substitution end	279	274			
Substitution beginning	11	10			
Substitution middle	18	9			
Total substitution	20	21			

C - Correct (%) **SS** - Some Sense (%) **W** - Wrong (%)

instance, in the second row, second column are the results of validation to the addition end model (1st pattern). The manual evaluation of the substitution model was not divided into patterns since, substitution patterns were equal between them. Some instances of the relations extracted are presented in tables 2, 3, 4 and 5.

The most interesting patterns are (ordered first by quantity and then by correctness):

1. Term(s) substitution at the end (1st and 2nd): 553 - 88%;
2. Term(s) addition at the end (1st): 76 - 82%;
3. Term(s) addition in the middle (4th and 5th): 30 - 96% 19 - 92%;
4. Term(s) deletion in the middle (4th and 5th): 39 - 90% 10 - 90%;
5. Term(s) deletion at the end (1st): 48 - 84%;
6. Term(s) addition at the beginning (3rd): 25 - 99%;
7. Term(s) deletion at the beginning (3rd): 21 - 78%;

Joining these seven patterns results in a total of 821 hyponymy relations.

It is possible to get better results by changing the search engine requirement. This can be done by restricting the relations added to those that obtained more than x hits. For instance, empirically we chose that addition end(2nd, 3rd), deletion beginning(2nd, 3rd) and total substitution(1st, 2nd), needed more than one entry to be assumed as a relation. But this value could be larger and thus the semantic relation would become stronger.

After discovering the most promising method (substitution end) using the search engine evaluation method, a sub-set of 1000 unique relations of this pattern was selected. From these 1000 relations we studied only 749 since the remaining relations contained orthographic errors. We then decided which were hyponyms and which were not. In the end about 40% were correct, 20% had some sense and 40% were wrong. We therefore estimate that among all the

27000 relations in the substitution end model, it is possible to extract about 12150 hyponymy relations.

Patterns with less than 20 relations have a very low frequency, thus do not represent the typical user search behavior and should not be used in hyponymy extraction.

5 Conclusions

Hyponymy extraction from logs reinforces the idea that human knowledge is organized like an ontology where parent-child relations exist [17]. Interestingly, users prefer to make a horizontal move in the ontology, as the frequency in the substitution model demonstrates. This movement allows us to extract co-hyponymy relations by connecting two hyponyms extracted in each of two queries.

The proposed user reformulation models can be used to extract hyponymy relations that can be added to ontologies. Taking into account the results obtained, we recommend the use of the 1st/2nd Substitution End and the 1st Addition End patterns in order to extract hyponymy relations. This method can improve ontologies by adding recent relations. The relations extracted tend to represent the knowledge of the community that more frequently uses the search engine.

Using Google alone and the best behavior patterns it was possible to extract 821 hyponymy relations. With the manual method it was estimated to be possible to extract about 12150 hyponymy relations with a single model.

In this paper we present a method to capture part of the web user knowledge, leading to an important update in the model of web query reformulation (Figure 1). With the capture of user knowledge it will be possible to develop web search engines that perform searches more efficiently.

In the future the study of other semantic relations (e.g. meronymy and holonymy), other logs[1] and other query reformulation patterns/models (e.g. analyzing the entire session or combining different sessions) should be performed.

6 Acknowledgments

We thank Paulo Gomes for his important support. A special thanks to everyone who participated in the experiments, without whom the study would not have been so complete. Part of this work was performed under a research grant given by CISUC/FCT.

References

1. Jansen, B.J.: Search log analysis: What it is, what's been done, how to do it. Library and Information Science Research **28** (2006) 407–432

[1] http://ist.psu.edu/faculty_pages/jjansen/academic/transaction_logs.html

2. Rieh, S.Y., Xie, H.I.: Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management **42** (January 01 2006) 751–768
3. Wang, P., Berry, M.W., Yang, Y.: Mining longitudinal web queries: Trends and patterns. J. Am. Soc. Inf. Sci. Technol. **54** (2003) 743–758
4. Chuang, S.L., Chien, L.F.: Enriching web taxonomies through subject categorization of query terms from search engine logs. Decision Support System **30** (2003)
5. Noriaki, K., Takeya, M., Miyoshi, H.: Semantic log analysis based on a user query behavior model. In: ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2003) 107
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th International Conference on Computational Linguistics, Nantes **S2K-92-09** (1992)
7. de Freitas, M.C.: Elaboração automática de ontologias de domínio: discussão e resultados. PhD thesis, Universidade Católica do Rio de Janeiro (2007)
8. Buitelaar, P., Cimiano, P.: Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Volume 167 Frontiers in Artificial Intelligence and Applications. IOS Press (March 2008)
9. Aires, R., Aluisio, S.: Como incrementar a qualidade dos resultados das maquinas de busca: da analise de logs a interaccao em portugues. Ciencia de Informacao **3** (2003) 5–16
10. Seco, N., Cardoso, N.: Detecting user sessions in the tumba! query log. Unpublished (March 2006)
11. He, D., Gker, A.: Detecting session boundaries from web user logs. In: 22nd Annual Colloquium on Information Retrieval Research. (2000)
12. Bruza, P., Dennis, S.: Query reformulation on the internet: Empirical data and the hyperindex search engine. In: RIA097 Conference Computer-Assisted Information Searching on Internet. (1997) 488–499
13. Saracevic, T.: The stratified model of information retrieval interaction: Extension and applications. In: 60th annual meeting of the American Society for Information Science. Volume 34. (1997) 313–327
14. Hancock-Beaulieu, M., Robertson, S., Nielsen, C.: Evaluation of online catalogues: An assessment of methods (bl research paper 78). London: The British Library Research and Development Department (1990)
15. Phippen, A., Sheppard, L., Furnell, S.: A practical evaluation of web analytics. Internet Research: Electronic Networking Applications and Policy **14** (2004) 284–293
16. Cimiano, P., Staab, S.: Learning by googling. SIGKDD Explor. Newsl. **6**(2) (2004) 24–33
17. Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior **8** (1969) 240–247