The Intention Behind Web Queries

Ricardo Baeza-Yates[1], Liliana Calderón-Benavides[2], and Cristina González-Caro[2]

[1] Yahoo! Research Barcelona
Ocata 1, 08003 Barcelona, Spain
ricado@baeza.cl
[2] Web Research Group
Universitat Pompeu Fabra, Passeig de Circumval-lació 8
08003 Barcelona, Spain
{liliana.calderon, cristina.gonzalez}@upf.edu

Abstract. The identification of the user's intention or interest through queries that they submit to a search engine can be very useful to offer them more adequate results. In this work we present a framework for the identification of user's interest in an automatic way, based on the analysis of query logs. This identification is made from two perspectives, the objectives or goals of a user and the categories in which these aims are situated. A manual classification of the queries was made in order to have a reference point and then we applied supervised and unsupervised learning techniques. The results obtained show that for a considerable amount of cases supervised learning is a good option, however through unsupervised learning we found relationships between users and behaviors that are not easy to detect just taking the query words. Also, through unsupervised learning we established that there are categories that we are not able to determine in contrast with other classes that were not considered but naturally appear after the clustering process. This allowed us to establish that the combination of supervised and unsupervised learning is a good alternative to find user's goals. From supervised learning we can identify the user interest given certain established goals and categories; on the other hand, with unsupervised learning we can validate the goals and categories used, refine them and select the most appropriate to the user's needs.

1 Introduction

Current Web search engines have been designed to offer resources to their users, but with the limitation that the goals or characteristics behind the queries made by them are not generally considered. Given that a query is the representation of a need, a set of factors, in most cases, are implicit within this representation. If we can discover these factors, they can be crucial in the information recommendation process. Techniques such as Web Usage Mining [1] cover the problem to improve the quality of information to users by analyzing Web log data. Particularly, Web Query Mining [2, 3], deals with the study of query logs from data registered in

F. Crestani, P. Ferragina, and M. Sanderson (Eds.): SPIRE 2006, LNCS 4209, pp. 98–109, 2006.

a search engine, with the purpose of discovering hidden information about the behavior of users of these kind of systems. Some work has been done to categorize the needs of users; for example, the categorization proposed by Broder [4] in which, according with the goal of the user, three classes are considered: Navigational, Informational and Transactional. Broder made a classification of queries through an user survey and manual classification of a query log. This work was later taken up by Rose and Levinson [5], who developed a framework for manual classification of search goals by extending the classes proposed by Broder. In their studies Broder, and Rose and Levinson showed that goals of queries can be identified manually. In Lee *et al.* [6] the work was focused on automatic identification of goals (navigational and informational) through the application of heuristics over clicks made by the users on the results offered by the search engine; in order to do it, they proposed two related features, the past user-click behavior and the anchor-link distribution. In the same context, works like Spereta *et al.* [7] tries to establish user profiles by using their search histories; Baeza-Yates *et al.* [2] discovered groups of related queries, through text clustering of documents clicked for the queries, allowing an improvement of the search process.

In general, the approaches try to make an approximation to the user from different perspectives. However, a model in which the user can be identified by using his/her goals hasn't been completely developed. We could then use this kind of information to understand and improve his/her information needs. Taking this into consideration, the main goal of this work is to develop a model for identification of the user's interests for a Web search engine, using the user interactions stored in the query log files of the system. The identification process is made from two perspectives, the first one is from the objectives or goals of each one of the users and the second is from the categories in which each of the objectives can be situated. To be able to measure precision and recall we manually classified more than 6,000 real queries, a reference set two orders of magnitude larger compared to the 50 CS related queries used by Lee *et al.* [6].

This paper is organized as follows. In Section 2 we describe user's goals and categories to which the user's queries can belong. A brief description of the used techniques to find user interest are presented in Section 3. In Section 4 we present the experimental design of this work. Finally, we present an analysis of the obtained results in Section 5 and conclude the paper in Section 6.

2 User's Goals and Categories

As a way to determine the motivations during an information search, we propose firstly, to find the user goals and secondly mapping these queries into categories. This information allow us to determine the path that a user follows when is searching for information on a web search engine.

2.1 User Goals

From the content of the queries we established three categories for the reasons or goals which motivate the user to make a search: *Informational, Not informational*

and *Ambiguous*. An informational query is one in which the user exhibits an interest to obtain information available in the Web, independently of the knowledge area of the resource retrieved. As not informational we categorize queries that find other resources or target an specific transaction (e.g. buy, download, reserve, etc.). Finally, ambiguous queries are those that their goal cannot be inferred directly from the query (in some cases because the user has an ambiguous interest). For Informational queries we could use a ranking biased towards text content while for Not Informational queries a better answer could be a few Web sites, good hub pages (many good links) or a price comparison portal.

2.2 Query Categories

A key point in the process of user interests identification is establishing the topic to which each submitted query belongs. The discovery of the kind of information requested allows us to identify it in a particular area of interest and relate it to specific characteristics of the area in which it is related (or in which he/she wants to be related).

Topics used to classify the queries are based on the general categories of the Open Directory Project, ODP[1] *(Arts, Games, Kids and Teens, Reference, Shopping, World, Business, Health, News, Society, Computers, Home, Recreation, Science, Sports)*. Apart of these general categories we considered three more which are: *Various* for those queries that from their content seems as belong to more than one category, *Other* for queries which can't be classified into one of the selected categories and *Sex* taking in account that a considerable amount of queries are related with this topic.

3 Selected Techniques

As a way to reach our purpose we selected two quite different models which, from the literature are available to categorize data and find hidden relationships among data. The selected models were Support Vector Machines (SVM) [8] and Probabilistic Latent Semantic Analysis (PLSA) [9].

3.1 Support Vector Machines Model

In this work, Support Vector Machines [8] have been used to build classification models for queries. We chose this classifier, given their proven effectiveness in different scenarios with a high feature dimensionality, including text classification [10]; considering that the queries were represented by the words of the pages selected by the users, this characteristic is quite useful. To solving the multiclass problem, we combine SVM with Error-Correcting Output Coding (ECOC) [11], which reduces the multiclass problem to a group of binary classification tasks and combine the binary classification results to predict multiclass labels. The RBF (Radial Basis Function) kernel was used to the SVM's setup, and we choose the kernel's parameters through a standard cross-validation process.

[1] Open Directory Project. http://dmoz.org

3.2 Probabilistic Latent Semantic Analysis

As we have commented, one of the main ideas that justifies the development of this work is to find the reasons which motivate the user to make a search in the Web. Considering this, and in accordance with different works such as Jin [12] and Lin [13], Probabilistic Latent Semantic Analysis (PLSA) [9] appear to be an efficient method of analyzing user interests.

Given that the starting point for PLSA is a statistical model which has been called Aspect Model [9], the implementation of this model used in this work was taken from PennAspect [14], a well tested software for information filtering and retrieval.

4 Experimental Design

Data Set. For this work we processed a log sample from the Chilean Web search engine TodoCL[2]. Thesample contains 6,042 queries having clicks in their answers. There are 22,190 clicks registered in the log, and these clicks are over 18,527 different URLs. Thus, in average users clicked 3.67 URLs per query.

Data Preprocessing. One of the most important ideas to exploit here is finding existing relationships in the data. In order to achieve this, each query was represented as a vector of terms that appeared in the documents giving an answer to the query (stop words were removed), $Q_i = w(t_1), w(t_2) \dots w(t_n)$, where w($tj$) is the associated weight of term j inside query Q_i. The classical TF-IDF weighting scheme was used to assign the weight to each query term and clicked page, replacing IDF by the number of clicks on each page (see [2]).

After that, a clustering process was applied over the data. We obtained query groups with similar characteristics, i.e. they belong to the same subject, are related with specific topics or describe the same situation using different terms. To do this, we used the simple K-means clustering method.

Manually Classified Data. To be able to evaluate the results of our automatic classification, we built a test set based on a team of people who performed a manual classification of the queries.

An important characteristic about the structure of these queries, as we mentioned before, is that they can be organized in clusters. This structure facilitated the manual classification process, by providing our team with information about the context of the query and at the same time, giving a global idea about the class to which each one of them belongs. This information is used to facilitate the human classifiers in the case that a query did not suggest a complete idea by itself. In any case, the task of a human classifier is to select the type of goal for a user and the category in which this goal can be situated. Considering the amount of queries and the different categories in which they can belong, the manual classification process is hard to do and subject to some subjectivity

[2] TodoCL. http://www.todocl.com

(and hence, errors). As a way to facilitate this process we created a software tool which offers to users the possibility to select the goals and the categories and save them in an organized and fast way.

Table 1. Manual classification of queries into goals and categories

Category	Inf	N-Inf	Amb	Total	Category	Inf	N-Inf	Amb	Total
Arts	102	23	29	154	Society	501	12	60	573
Games	11	26	8	45	Home	50	35	41	126
Education	232	29	23	284	Recreation	789	489	142	1,420
Reference	107	85	26	218	Science	129	7	9	145
Shopping	55	29	39	123	Sports	31	11	5	47
World	46	6	15	67	Computers	174	208	86	468
News	78	5	1	84	Sex	37	178	33	248
Business	960	107	93	1,160	Others	16	9	33	58
Health	171	21	40	232	Various	224	27	339	590

As we previously described, we established three categories to which the goals can belong: Informational (Inf), Not Informational (N-Inf) and Ambiguous (Amb). On the other hand, we established eighteen topics to classify the same queries. After the manual classification process of queries into goals and categories, the obtained amount are presented in the table 1.

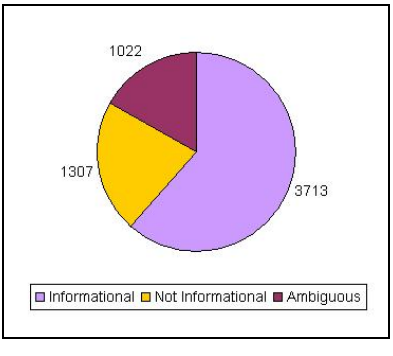

Fig. 1. Distribution of Queries into Goals

Manual classification of queries into goals. The figure 1 presents the amount of queries which were labeled by our team as Informational, Not Informational and Ambiguous. The goal with the higher number of queries was Informational, this happens because we considered as the kind of queries which not talking directly about an object (such as mp3 file, photo, among others), the name of an artist or the purchase or sale of a product or service. However the goals categories considered to label a query are different between this work and the work realized by Rose and Levinson [5], we agreed on the proportion of Informational queries, being this higher than the others.

Manual classification of queries into categories. A graphical representation of the manual classification of queries into categories is presented in the figure 2. The categories with higher amount of queries are Entertainment and Business, which confirm the search behavior of people that have been well described by Spink and Jansen in their works [15, 16].

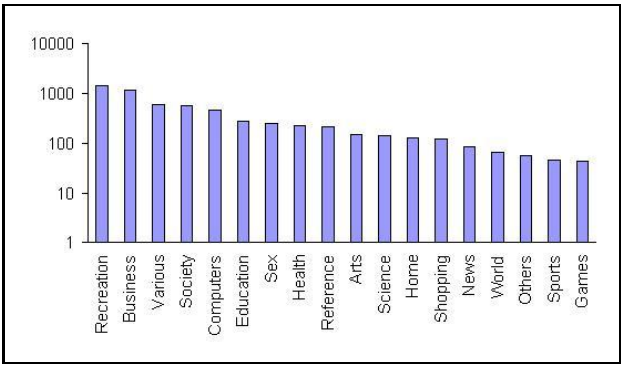

Fig. 2. Distribution of Queries into Categories

The figure 3 present the distribution of the percentage of categories into the different goals. The queries grouped as Informational goal belong to categories such as Business, Education, Science or News in which people are searching for resources answering in many of the cases to a specific information need. On the other hand, queries grouped as Not Informational belong to categories such as Recreation, Sex or Games in which the intention is, in most of the cases to visit a place to find one of this kind of sources. Finally, the queries grouped as Ambiguous are more present in the Various and Other categories due that it is not clear what the user wants and hence are quite difficult to classify in one of the other.

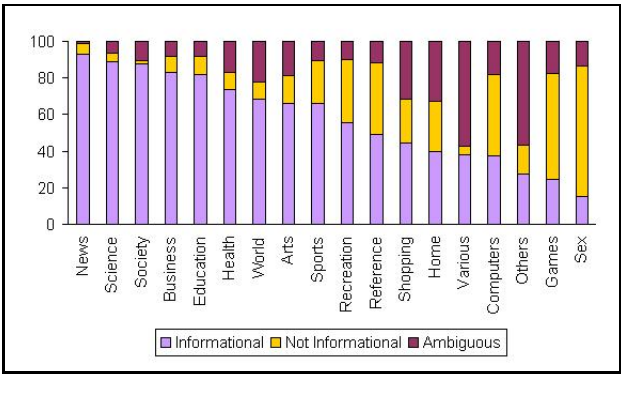

Fig. 3. Percentage distribution of queries into Goals and Categories

Performance. For the training phase, PLSA took, on average, four hours to build a model and calculate the different probabilities of each query and the words belonging to each latent class. As we mentioned before, we determined three goals and selected eighteen categories to which the user interest can be situated and their queries can be classified. These quantities were used to generate the query groups (in section 5.2 we will comment about this fact).

To build the models and make the predictions for categories, SVM spent about two hours. For the case of goals, considering the low amount of labels involved, this model took about fifty minutes.

The different algorithms were run on a Pentium III computer with 1.28 GBs of RAM, under a Linux OS.

5 Analysis of Results

5.1 Supervised Learning

After the manual classification of the queries was made, part of these labeled data was used like input to train an automatic classifier.

The obtained results in the classification process with supervised learning were good. From the labeled examples by our team of editors, quite suitable models for each goal and category were constructed. Although, not in all the cases the predictions agree with the human judgments, the prediction in most cases is related with the subject of the query, showing therefore the ambiguous nature of some queries that can be located in different goals or categories (our Ambiguous class). Nevertheless, in the case of the categories, the idea was, as far as possible, to assign a determined category to each query, the category "Various" was used as minimum as possible in the manual classification.

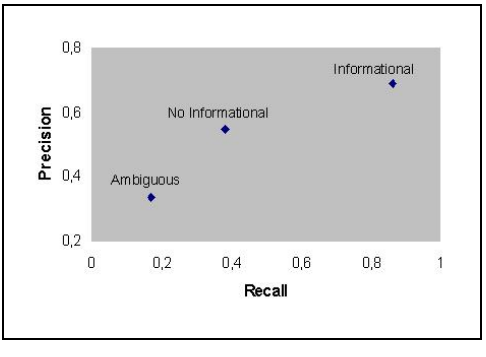

Fig. 4. Recall-Precision Graph of Goal's Automatic Classification

With respect to the user's information goals the results are good, the precision is over 50% for two of the goals. The best results are obtained with the Informational goal, its precision is high and its recall is almost perfect, see figure 4. This is due that the pages selected for the queries that belong to this goal are more

homogeneous, unlike those that belong to "Not Informational" or "Ambiguous", where the nature of the pages selected by the users is heterogeneous, since the users do not have a very precise idea of what they wish to find.

For the categories, the results are good in general; nevertheless, for some categories the results are better than for others. In the figure 5 we can see a sample of the most representative categories:

The categories that show better precision are those that have greater popularity, that is to say, those related to subjects that the people consult most frequently, like for example: Recreation. Since most of pages of this type of subjects handle a moderately similar vocabulary, the queries are more identifiable. Categories as "Technology and Computers" have, relatively, a specialized vocabulary, which allows to identify more accurately the related queries. Another particular case is the category "Sex", where the users do not change the words used to make their queries, most of these are built using the same words, without counting that, even though the users use different words to describe their queries, the pages which they choose as answer does not change, so the queries are repeated from the same words again.

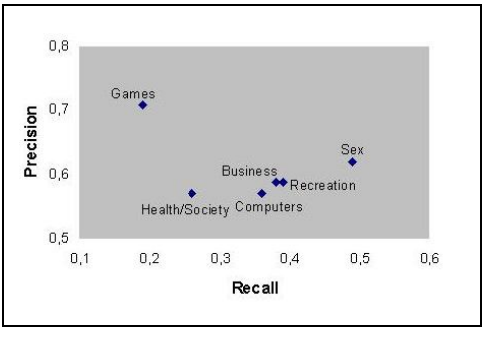

Fig. 5. Recall-Precision Graph of Categories's Automatic Classification

When we analyzed the relationship between goals and categories, we can observe the coherence that exists between the informational objectives of the users and the categories in which their queries are located. For the Informational goal, the greater distribution of queries are in the categories: Business, Recreation, Society, Education and News. Whereas the categories with smaller concentration of queries in this goal are: Games, Shopping, Home, Sex and Others, these last categories, suggest different motivations for the user that are not related to obtain information, for example Games, where the users are more interested in downloading software and resources. A particular case is the category "News", that it concentrates all its queries in this goal, these queries don't belong to other goals. Similarly, the goal "No Informational" shows enough coherence with the categories of the queries that belong to her. In this case, the categories with greater concentration of queries are Sex, Recreation and Technology and Computers, and those of smaller concentration of queries are Society, Health, Science

and News (as they do not have queries in this goal). Finally, "Ambiguous", that it's a goal that as its name indicates it, is not related to some category in individual, the greater concentration of queries is in the category "Various", which is logical, given the nature of the category and the goal.

5.2 Unsupervised Learning

Categories. To determine the relationships, at the categories level, between different queries used in this work, we considered topics (described in section 2.2), to which the queries can belong, as the variables that make that a user submit a specific query.

Before analyzing the obtained results with PLSA, and having in mind that this work is focussed on user interest, is important to highlight that in a common process of information search we are exposed, first, to the lack of precision between the transformation of a mental information need to a set of key terms that correctly describes this need, and second, to the lack of accuracy of search engines to provide an answer including aspects such as subjectivity or the context of the searching task. However, taking advantage of the results offered by PLSA model, and it's capability to provide a probabilistic information about the degree of membership of each query to each generated cluster, we can make an analysis of the composition of each cluster content. This information offer us the possibility to discover direct and not direct relationships between queries and topics (i. e. to which categories a query can belong), and from this information we can determine what is the user interest.

One of the most important aspects to highlight here is that although the amount of clusters, used to make the clustering process for categories identification of user interests, was taken from the ODP categorization, the obtained results from PLSA shows a hard grouping of queries around some of the categories such as Sex, Entertainment, Business, References or Health. However, the model could not create significant groups for categories like Arts, Sports, Science or Games. This happens not only because the amount of queries is very low, but also because they are mixed with other unrelated queries. This information was used from two point of views:

- Ratify that most of the selected categories used in the manual classification are clearly defined. However, there are other categories that have overlapped content and are difficult to determine. In contrast to these facts, other possible categories that we did not consider appeared, such as cars and law.
- From this information we can identify existent relationships between queries. The table 2 shows an example of these relationships. In this table we have a sample of queries grouped in cluster 6, which was labeled as Recreation or Entertainment. By observing the probabilities values (Prob1, Prob2, Prob3) of each one of these queries belonging to each cluster, the highest values are for clusters labeled as Business (cluster 7) and Sex (cluster 11).

In general terms, we can say that queries that were grouped in the Sex category, have a high probability to belong to entertainment, which is absolutely

Table 2. Queries with three highest probabilities in the Recreation cluster

QId	Query	Prob1	CId	Prob2	CId	Prob3	CId
4197	los jaivas main works	1.76E-03	6	1.99E-09	0	4.30E-12	7
243	ricardo arjona spanish songs	1.51E-03	6	2.01E-08	7	2.12E-42	11
5759	madonna erotic	1.50E-03	6	1.83E-08	7	2.20E-42	11
1917	porto seguro cd	1.50E-03	6	2.68E-08	7	1.29E-43	11
5378	rata blanca songs	1.50E-03	6	2.69E-08	7	8.84E-43	11

coherent; on the other hand, these same queries can be considered as belonging to business category due to that the content of pages answering sex or entertainment queries have terms related with payments or selling this kind of services.

A particular case was presented by the cluster which grouped queries related to health. About 70% of queries belonging to this cluster made reference to drugs, diseases or treatment of diseases. The reason for this case is that the medical vocabulary and the terms used to make this kind of queries are too specific, and is quite rare to find problems of synonimy or polysemy. The table 3 shows the five queries with highest probability in this cluster.

Table 3. Queries with highest probabilities in the Health cluster

IdQuery	Query	Probability
1831	electroconvulsive therapy	1.75E-03
2215	nasal polyps	1.53E-03
3507	dental hygienist	1.51E-03
3156	hepatitis	1.41E-03
5023	viagra	1.03E-03

Goals. Through PLSA we found that approximately 73% of the 2,168 queries grouped as ambiguous, belong to categories such as Sex or Entertainment. It is important to note that none of the queries, labeled in the manual classification as Health, is part of an ambiguous goal, as a person usually has in mind the name of an specific illness or drug.

From the 2,719 queries grouped as Informational, about 76% are related with References, Education, Health, Computers, Society and Home. Finally, from the 1,155 queries grouped as not informational, near to 70% were labeled (in the manual classification) as Computers, Entertainment, Society and Sex. The main difference between queries that belong to the ambiguous cluster and the not informational queries is that the second makes direct reference to a photos of famous artists or models, parts of computers and software downloads, and songs, among others.

6 Conclusions

In this work we have presented a first step to identify user's interests in a Web search engine based in a query log. An analysis was made from two perspectives:

the user's informational objectives and the categories in which the queries within these objectives can be located. In order to identify these interests, different techniques were used, initially a manual classification, whose objective was to make a recognition from the human judgments of the distribution of goals and categories that could have the queries to classify. Later, we carried out an automatic identification of these interests using supervised and unsupervised learning.

The supervised analysis allows us to establish that the user interests are identifiable using a particular representation of queries, together with the automatic classifier. This was a good combination, since representing the queries by the terms of the documents that gave good answers to them, reduces the problem of the low number of words that the users use to make their queries (and hence the sparsity of the query space), and additionally because the pages that belong to a same category share a similar vocabulary that allowed us to make a better classification.

From the unsupervised perspective, user needs related with entertainment, sex or business were very well detected and important relationships between these categories were reflected. Most of the queries that were grouped in one of these three categories can belong to another categories. On the other hand, not for all the proposed categories exist a strong way to determine users' needs. This happen because the terms used to summit a query and the content of the pages in an answer to this query can be used to describe different topics. From the eighteen proposed categories, just eleven of them were completely recognized. In the opposite, two new and well defined categories appeared in the clustering process, they were cars and law. This suggests to make a revision of the selected ODP categories, avoiding overlapping of information.

The bottom line is that for the Informational class and some categories we got over 70% precision and very good recall. This can be easily improved by trying other query representations, other classification techniques, etc.

Acknowledgements

The authors wish to thank Mari-Carmen Marcos for helpful comments and suggestions in the classification of queries. The authors are grateful to the Information Technologies Research Group from the University Autónoma of Bucaramanga for help in the manual classification process of queries. This work was partially supported by the Alpha Project AML/B7-311/97/0666/II-0291-FA.

References

1. Mobasher, B. In: Practical Handbook of Internet Computing. CRC Press (2005)
2. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Current Trends in Database Technology - EDBT, Springer-Verlag GmbH (2004) 588–596
3. Baeza-Yates, R.: Applications of web query mining. In: ECIR 2005. Volume 3408., Lecture Notes in Computer Science (2005)

4. Broder, A.: A taxonomy of web search. SIGIR Forum **36** (2002) 3–10
5. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: International conference on WWW, ACM Press (2004) 13–19
6. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: International conference on WWW, ACM Press (2005) 391–400
7. Speretta, M., Gauch, S.: Personalizing search based on user search history (2004)
8. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2** (1998) 121–167
9. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, Stockholm (1999)
10. Basu, A., Watters, C., Shepherd, M.: Support vector machines for text categorization. In: International Conference on System Sciences, Washington, DC, USA, IEEE Computer Society (2003) 103.3
11. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of AI Research **2** (1995) 263–286
12. Jin, X., Zhou, Y., Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis. In: Knowledge discovery and data mining, New York, USA, ACM Press (2004) 197–205
13. Lin, C., Xue, G.R., Zeng, H.J., Yu, Y.: Using probabilistic latent semantic analysis for personalized web search. In: Web Technologies Research and Development, Berlin Heidelberg, Springer-Verlag GmbH (2005) 707–717
14. Schein, A., Popescul, A., Ungar, L.: Pennaspect: A two-way aspect model implementation. Technical report, (Department of Computer and Information Science, The University of Pennsylvania)
15. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the web: the public and their queries. Journal of the American Society for Information Science and Technology **52** (2001) 226–234
16. Jansen, B.J., Spink, A.: An analysis of web searching by european alltheweb.com users. Information Processing and Management: an International Journal **41** (2005) 361–381