Classifying queries into directories: an approach based on click-through data

Marcelo Mendoza

Department of Computer Science
Universidad de Valparaiso
marcelo.mendoza@uv.cl

Abstract. In this paper we claim that a query classification method could be useful to discover relationships among queries and the concepts behind a Web directory. This assumption is based on the fact that Web directories are hierarchical structures of knowledge where each node represents a well defined concept. Thus, if we can classify a query into a node, we enrich the node description. Exploring the idea of processing the user's click data to classify queries into Web directories, we build a vectorial representation of query sessions based on click-through data. Using a nearest neighbor classifier, we classify queries into categories following a top-down approach. Our experimental results suggest that the proposed method is effective and it could be useful to improve the precision of the search engine answer lists.

1 Introduction

Query mining is an important topic for the web community and the search for non trivial patterns in the query log data could be useful to improve the precision of the answer lists recommended by the search engines, as shown in [2]. Due to the fact that the answer lists could be compounded by millions of pages, it is neccesary to provide structures to the information presented to the users. Among other alternatives, the most relevant information structure used to present results to the users is the web directory.

Web directories are hierarchies of classes which cluster documents covering related topics [1]. Directories are compounded by nodes where each node represents a category where documents are classified. The structure of a directory is as follows: the root category represents the *all* node. The all node means a complete corpus of knowledge. Thus, all queries and documents are relevant to the root category. Each category shows a list of documents related to the category subject. Traditionally, documents are manually classified by human editors.

The categories are organized in a child/parent relationship. The child/parent relationship represents a generalization/specialization relation among the concepts represented by the categories. According to Chakrabarti [3] we understand this kind of relationship as an *inheritance* relationship: If a category c_0 is the parent of a category c_1, any web item that belongs to c_1 also belongs to c_0. If we understand a parent/child relation as an "inheritance" relationship, any web

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 182-189, 2008.

item that belongs to a descendant of a given category represents a *specialization* of its meaning. Conversely, any web item is related to the meaning of the parent categories and the relationship among them represents a *generalization* of its meaning.

In this paper we claim that a query classification method could be useful to discover relationships among queries and the concepts behind a Web directory. This assumption is based on the fact that Web directories are hierarchical structures of knowledge where each node represents a well defined concept. The identification of relationships among queries and nodes (concepts) could be useful for the automatic maintenance of the Web directory as shown in [4]. Intuitively, if we can classify a query into a node, we enrich the node description, adding queries and the documents selected in their sessions to the list of recommended items.

A good quality query classification method should consider a minimal consistency among the classification rules and the structure of the taxonomy in its design. To formalize this idea, we define a principle from the consistency of a query classification: Let c be a category in a taxonomy τ and q be a query *semantically* related with c. If q is classified into c, the classification is consistent with τ only if q is semantically related to the parents of c.

Frequently a query classification schema is worked out following a flat approach. A flat approach classifies the query into the nearest category using a distance function, without any constraint related to the structure of the taxonomy. The main problem of flat models is the possible violation of the consistency principle. In general a flat model does not guarantee the consistency principle.

Contributions. In this paper we introduce a method for query classification according to the consistency principle. Our aim is to explore the idea of processing the user's click data to classify queries into a Web directory building a vectorial representation of query sessions using click-through data. Using a nearest neighbor classifier based on distance functions, we classify queries into categories following a top-down approach. Experimental results show in this paper that it is possible to improve the precision of the search engine recommendations following the proposed method.

2 Related work

Chakrabarti [3] proposes a Bayesian approach to classify queries into subject taxonomies. Based on the construction of training data, the queries are classified following a breadth first search strategy starting from the root category and descending one level on each iteration. The main limitation of the method is the following: queries are always classified into leaves because the leaves maximize the probability of the tree path.

In the *KDDCUP'05*, the proposed competition was focused on the classification of queries into a given subject taxonomy. To do that, the competition organizers provided a small training data set composed by a list of queries and

their category labels. Most of the papers were based on classifiers which learn under supervised techniques. The winning paper was written by Shen *et al.* [8] and applies a two phase framework to classify a set of queries into a subject taxonomy. Using a machine learning approach, they collected data from the web for training synonym based classifiers that map a query to each related category. In the second phase, the queries were formulated to a search engine. Using the labels and the text of the retrieved pages, the queries were enriched in their descriptions. Finally, the queries were classified into the subject taxonomy using the classifiers through a consensus function. The main limitations of the proposed method are the dependency of the classification to the quality of the training data, the human effort involved in the training data construction and the semi-automatic nature of the approach which limit the scale of the method's applications.

Vogel *et al.* [9] classify queries into subject taxonomies using a semi-automatic approach. First they post the query to the Google directory [6] which scans the *Open Directory* [5] for occurrences of the query within the Open Directory categories. Then the top-100 documents are retrieved from Google formulating the query to the search engine. Using this document collection an ordered list with the categories assigned to the 100 retrieved documents is built. Using a semi-automatic category mapping between the web categories and a subject category the method identifies a set of the closest topics to each query. Unfortunately, the method is limited to the quality of the Google classifier that identifies the closest categories in the Open directory. Also, it is limited to the quality of the answer lists retrieved from Google. Finally, the semi-automatic nature of the approach limits the scale of the method.

3 The classifier

Search engines show their recommended documents to queries as lists of items where each item is formed by the document URL, the title and a *snippet* (excerpt). Intuitively, if the snippet, the title and/or the document URL are semantically related to the user's intention, then the user will select the document.

We will consider the terms of the snippets, the title and the document URL to build a term-weighted vectorial representation. In order to do this, we measure the relevance of each excerpt considering another variable useful for our representation: the time spent in each document visit. We will use the following assumption: the more relevant the document is, the longer the user will spend visiting it. Finally, we will also include query terms. For our representation the query is another selected document where its meaning is expressed by the query terms.

Now we will formalize the representation. Given a query session s, let q be a query formulated into s, and let U_s be the set of documents selected in s. For a document u in U_s, let $Tf_{t,q}$ and $Tf_{t,u}$ be the number of occurrences of the term t in the query q and in the document snippet, title and URL of u, respectively. We build our representation from the documents in U_s and the query q considering

a $tf - idf$ scheme proportional to the time t_u spent in each document selected, and normalized by the total time t_s in the session s.

Let $\overrightarrow{v_s}$ be the term vector for a query session s, where $v_s[i]$ is the i-th component of a vector associated to the i-th term of the vocabulary. The i-th component $v_s[i]$ of the vector is defined as follows:

$$v_s[i] = \left(0,5 + 0,5\frac{Tf_{i,q}}{max\ Tf_q}\right) \times \log\frac{N_Q}{n_{i,Q}} \times \frac{1}{\mid U_s \mid} + \sum_{u \in U_s} \frac{1}{\mid U_S \mid} \times \frac{Tf_{i,u}}{max\ Tf_u} \times \log\frac{N_U}{n_{i,u}} \times \frac{t_u}{t_s}, \quad (1)$$

where Q is the query collection (the set of queries formulated and registered in the logs), N_Q is the number of queries in Q, $n_{i,Q}$ is the relative frequency of the i-th term in Q (the number of queries in Q where the i-th term appears), U is the document collection (the set of selected documents registered in the logs), N_U is the number of documents in U, and finally $n_{i,u}$ is the relative frequency of the i-th term in the document set U (the number of documents in U where the i-th term appears).

The first half of the equation represents the weight of the i-th term considering the query as an information source. We use the schema proposed by Salton and Buckley [7] in order to avoid a sparse query term vector. The second half of the equation is a sum of the weights of the i-th term for each selected document in the session. Each weight is normalized with the time spent by the user in the visit. Intuitively, a term will be more relevant for the representation if the term has a significant number of occurrences in the document (*tf* factor) and the document is relevant for the user (time factor). We use the *idf* factor in order to incorporate the effect of the distribution of the term in the entire collection.

To calculate this representation, we retrieve the snippet, the title and the document URL for each pair query-document registered in the log. Terms are processed in order to eliminate *stopwords*. Visit times are also calculated from the user's click data.

Similarly, for a category c, we obtain a term vector representation $\overrightarrow{v_c}$, by aggregating the text of every snippet, title and document URL that appears in the list of recommended documents of c.

Each query will be classified following a top-down approach. First, we will determine the closest centroid to the query considering all the centroids at the first level of the tree in the concept taxonomy. Then we repeat the process for each level of the taxonomy while the distance between the query and the closest centroid will be less than the distance at the previous level.

Now, we formalize the method. Let q be a query in a collection C of queries registered in the logs and let $c_{i,j}$ be the i-th category in the j-th level of a web taxonomy τ. For the zero level of the web taxonomy, the nearest category to the query is determined. Let $c_{*,0}$ be the closest category to q at the zero level of τ and $\Gamma(c_{*,0})$ be the descendant set of $c_{*,0}$ at the following level. In the next iteration of the method the classifier calculates the distances between q and

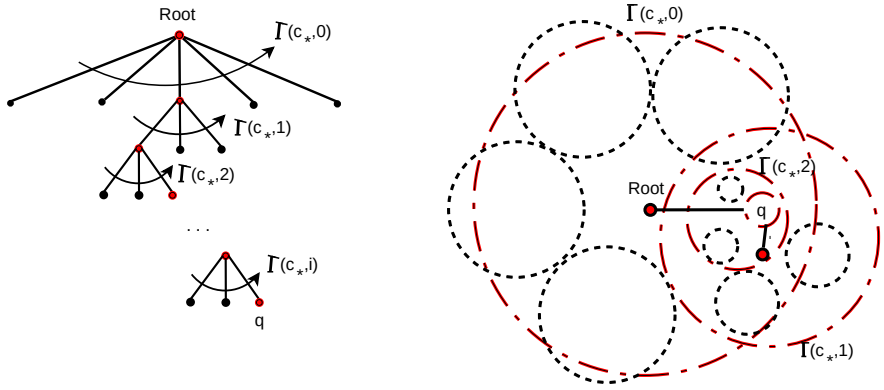

Fig. 1. The top-down classification schema proposed

each category in $\Gamma(c_{*,0})$, as is shown in Figure 1. Then the nearest category in $\Gamma(c_{*,0})$ is determined. Let $d_{min}(q, c_{*,1})$ be the distance between q and the closest category in $\Gamma(c_{*,0})$ and $d_{min}(q, c_{*,0})$ be the distance between q an the closest category at the zero level of τ. If $d_{min}(q, c_{*,1}) < d_{min}(q, c_{*,0})$ then q is classified in the closest category of the first level of τ. Then, in the next iteration of the method, the classifier calculates the distances between q and each category in $\Gamma(c_{*,1})$, repeting the process one level down. Otherwise, the method stops.

4 Experimental results

In order to evaluate the proposed method, the following experiments will be carried out. First of all, we retrieve a log file of 6 months which contain 127,642 queries over 245,170 sessions from TodoCL, a Chilean search engine. There are 617,796 selections registered in the log and these selections are over 238,457 different URLs. Thus in average users clicked 4.84 URLs per query.

We intend to illustrate with examples that the classification method has the ability to identify concepts related to the query. To do this, we randomly select 30 queries from the total. On the 30 queries, we will carry out experiments that allow us to evaluate the appropriateness of the classification and its usefulness for the user. Table 1 shows the 30 queries considered in our experiments, the taxonomy nodes where they are classified and the distance between the vectorial representations of the initial query and the category.

The plus sign indicates specialization in the taxonomy. For example, the *art+ music + midi* node shows that the query about Chilean music history was classified into the *art* topic, *music* section and into the *midi* subsection. As we can see in Table 1, each selected query is related to a well defined concept that describes one possible meaning. None of the classified queries was a false positive.

Now we will evaluate the quality of the answer lists retrieved by the search engine ranking method, the method based on the top-down classifier, and a

Query	taxonomy node	distance
Romane ratings	travels tourism	0,99
interactive museum Mirador	education	0,988
yoga	health	0,978
work environment complaints	government	0,974
Patricio Del Canto	arts + museums and cultural centers	0,973
Sinergia	arts + music + bands artists	0,973
Francisco Moya	arts + galleries	0,969
Metalcon foundation blocks	economy and businesses + industries + forests	0,967
jewelry lessons	arts	0,963
signs publication	arts + graphic arts	0,952
rolls of grass	home + gardening	0,948
clothing projects	economy and businesses + textiles + clothes	0,947
southern road	tourism travels	0,942
financial concepts	economy and businesses + finances	0,92
law 14,908	society + family	0,917
Metal furniture	home	0,895
shows	arts and entertainment + audiovisual production	0,877
X region companies	economy and businesses	0,871
educational evaluation issues	education	0,87
houses for sale in Iquique	regions + geographic zones	0,868
companies in Chile	guides directories	0,855
transpersonal psychology	health + psychology	0,85
hosting	guides directories + portals	0,822
furniture sales	home	0,818
Antofagasta Clinic	health + clinics and hospitals	0,815
satellite telephony	economy and businesses + telecommunications	0,815
PSU results	education + university selection test	0,814
Chilean music history	arts + music + midi	0,796
properties	economy and businesses + estate agencies market	0,761
sanitary engineering	economy and businesses + environment	0,756

Table 1. Queries selected for the evaluation of the method of query classification in directories sorted by distance.

standard method based on a flat classifier proposed in the literature [4]. In order to do this, we have considered the first 10 documents recommended by the search engine for each one of the 30 queries, the first ten documents recommended by the web directory when the closest category is determined using the flat classifier, and the first 10 documents recommended by the web directory when the query is classified using the proposed method. The document quality has been evaluated by a group of experts using the same relevance criteria as in the previous experiment (0-4, from lower to higher relevance). The precision for every ranking and every query is obtained, according to the position. Finally, the average precision is calculated over the set of documents recommended by position. Results are shown in Figure 2.

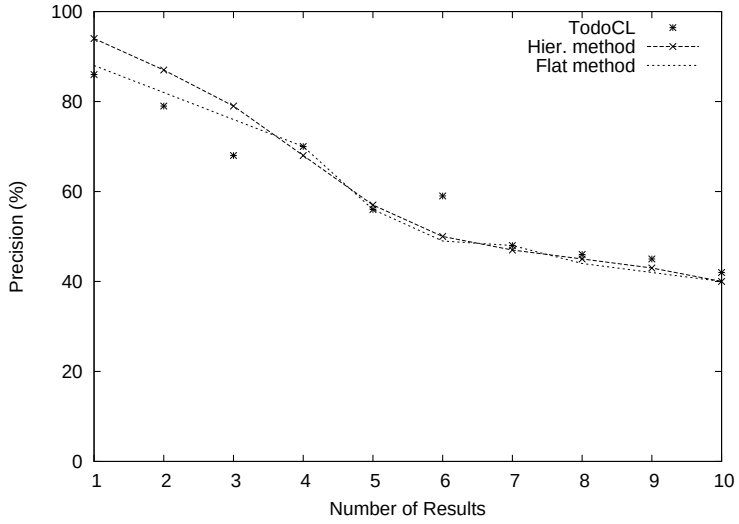

Fig. 2. Average precision of the retrieved documents for the methods based on classifiers and for the search engine ranking.

In Figure 2 we can observe that the evaluated methods are good quality rankings, especially for the first 5 recommended documents. The recommendation methods based on hierarchical classification and flat classification perform in a better way for the first 5 positions than the original ranking. This means that the new ranking is better than the original ranking, meaning that this is a good quality classification scheme. However, the ranking loses precision compared to the one of the search engine, if we consider the last 5 positions. This is due to the fact that many of the evaluated queries are classified into taxonomy nodes where less than 10 documents are recommended. In these cases, since there is no recommendation, the associated precision equals 0, which severely disqualifies the methods based on classifiers. Fortunately, none of the queries are classified into a node with less than 5 recommended documents. Therefore, a

fair comparison of the methods should be limited to the first 5 positions where, as we have seen, the proposed method is favorably compared with the original ranking and with the flat classifier.

5 Conclusion

We can conclude that our query classification method allows us to identify concepts semantically related to the queries. The proposed method also allows us to improve the precision of the retrieved documents over the first 5 positions of the answer lists. Compared with the search engine ranking method and compared with the method based on a flat classifier, the proposed method provides better results regarding the precision of the answer lists. As future work we propose to do a more conclusive analysis of the results, considering more intensive experiments and incorporating to the experimentation performance metrics such as ROC or F1.

One of the biggest limitations of the proposed method lies in the fact that it depends strongly on the taxonomy quality. This limitation is related to the fact that since the directory is manually maintained, it is limited in enlargement and freshness and in general the coverage of directories is low. Fortunately some authors are addressing this problem [4], providing methods for the automatic maintenance of web directories.

Acknowledgements

Marcelo Mendoza was supported by DIPUV 52/2007 project from Universidad de Valparaíso, Chile.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, ACM Press, New York, 1999.
2. R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Improving search engines by query clustering. *JASIST*, 58(12):1793–1804, 2007.
3. S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.
4. A. Cid, C. Hurtado, and M. Mendoza. Automatic maintenance of web directories using click-through data. In *Proceedings of ICDE workshops, Atlanta, Georgia, USA, April 2-4*, IEEE Computer Society, page 43, 2006.
5. DMOZ. *Open Directory Project*. `http://dmoz.org/`.
6. Google. *Google search engine*. `http://www.google.com/`.
7. G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
8. D. Shen, R. Pan, J. Sun, J. Junfeng, K. Wu, J. Yin, and Q. Yang. Q2caust: Our winning solution to query classification in kddcup 2005. *SIGKDD Explorations*, 7(2):100–110, 2005.
9. D. Vogel, S. Bickel, P. Haider, R. Schimpfky, P. Siemen, S. Bridges, and T. Scheffer. Classifying search engine queries using the web as background knowledge. *SIGKDD Explorations*, 7(2):117–122, 2005.