Building decision trees to identify the intent of a user query

Marcelo Mendoza[1,3] and Juan Zamora[2]

[1] Yahoo! Research, Santiago, Chile
[2] Department of Informatics, Universidad Técnica Federico Santa María, Chile
[3] Computer Science Department, Universidad de Valparaíso, Chile

Abstract. In this work we explore the use of decision trees to identify the intent of a user query, based on informational, navigational, and transactional categorization. They are based on decision trees, using the C4.5 implementation. The classifier will be built from a query data set larger than any previously used, allowing the conclusions to have a greater reach. Unlike previous works, we will explore characteristics that have not been evaluated before (e.g. PageRank) combined with characteristics based on text and/or click-through data. The results obtained are very precise and the decision tree obtained allows us to illustrate relations among the variables used for classification determining which of these variables are more useful in the classification process.

1 Introduction

At present, the Web is the largest and most diverse document database in the world. To access the content, Web users use search engines, where they formulate their queries. Then, the most relevant sites and pages are displayed for the user as a list of answers ordered by relevance to the query. Given the vastness of the Web and the little information on which a search engine has to base the query, the answers lists tend to be imprecise.

Broder [2] suggested that a way to improve the precision of the answers lists was to distinguish between the query and the needs of the user that formulates the query (the intent of a user query). That way it would be possible to use ranking algorithms that are adjusted to the type of user need.

Broder identified three types of needs from the queries. The first type is *Informational*, in which users search for information available in the content of the sites / pages. The most common form of interaction with this type of content is reading. Second is the *Navigational* need, where users search for a specific site whose URL they do not remember. The third type of need is *Transactional*, in which the users search for a page / site to make some kind of transaction such as download a file, make plan reservations, buy / sell, etc.

Later, numerous works concentrated on identifying characteristics of the queries that allow them to be categorized with the Broder taxonomy. These classifiers seek to automatically classify the queries in the categories described

earlier. There has been a wide range of results, and in general, there are no conclusive results. This is largely because these classifiers have been evaluated using very small data sets or under experimental conditions that make it impossible to generalize the results obtained.

2 Related Work

Once Broder proposed the web search taxonomy, defining the informational, navigational, and transactional categories, Rose and Levinson [10] extended Broder's categories. They identified types of frequent interactions between users and the recommended sites / pages, depending on the type of query posed.

Following this line of investigation, Kang and Kim [4] proposed characterizing the queries based on the distribution of the terms they contain. Based on a set of queries pulled from the TREC[4] collection and classified by experts into the Broder categories, they obtained a collection of terms frequently used to pose navigational or informational queries. Using mutual information criteria between the query terms and the titles and snippets from the selected pages / sites in the query sessions, they were able to automatically classify the queries with nearly 80% precision, on an evaluation data set of 200 queries.

Later, Lee *et al.* [5] proposed identifying the type of query by observing the levels of bias in click distributions in the query sessions being classified. Intuitively, an informational query should have a distribution with more clicks than a navigational query, where it is presumed that user preferences will generally be concentrated on just one site. From this idea, they proposed a classifier that reached near 50% precision on an evaluation data set of 50 queries.

A similar strategy was proposed by Liu *et al.* [6] who introduced two measurements based on click-through data to characterize queries based on their session logs in the search engine: `nRS` (number of sessions in which clicks are registered before a given position n) and `nCS` (number of sessions registering less than n clicks). Using a decision tree they were able to reach close to 80% precision on 400 manually classified queries. Later, Baeza-Yates *et al.* [1] proposed using a query vector representation based on text and click-through data. Using techniques such as Support Vector Machines (SVMs) and Probabilistic Latent Semantic Analysis (PLSA), they reached almost 50% precision on a dataset of 6,000 queries semi-automatically classified into the Broder categories (the vector representations were clustered and then the clusters were labeled into the categories).

Recently, using classifiers based on characteristics extracted from the query session registry, such as the number of query terms, number and distribution of clicks, among others, Jansen *et al.* [3] achieved nearly 74% precision on an evaluation data set of 400 queries manually classified by a group of experts. Finally, on the same data set used in the Baeza-Yates *et al.* [1] experiments, Mendoza and Baeza-Yates [7] showed that the characteristics based on text or

[4] Text Retrieval Conference co-sponsored by NIST. Dataset available on *http://trec.nist.gov/data.html*

on combinations of text and click-through data differentiated more queries in the Broder taxonomy than those based just on click-through data.

3 The classifier

3.1 Data set

Based on a query log file provided by AOL[5], consisting of a three-month period in 2006, which contains 594,564 queries associated to 765,292 query sessions, registering 1,124,664 clicks over 374,349 selected pages / sites, we will analyze the characteristics that will be most useful for the classification process. Experts from the *Universidad Tecnica Federico Santa Maria* of Chile and the *Universitat Pompeu Fabra* of Spain have collaborated in the manual classification of 2,000 queries randomly pulled from the AOL log. The experts completed a survey similar to that used by Broder in his first manual categorization experiment [2]. The final set of queries used for the analysis was compiled from those queries where the answers of the experts were in agreement. As a result, 1,953 queries were labeled by consensus, discarding only 2% of the initial data set. The queries categorized by consensus were distributed from greater to lesser proportion in the informational, navigational, and transactional classes, with 52%, 33%, and 15% over the complete data set, respectively.

3.2 Feature Analysis

Based on the characteristics analyses by Lee *et al.*[5], by Liu *et al.*[6] and by Mendoza and Baeza-Yates [7], we will use the following characteristics for the classification process, selected according to their discriminatory capacity:

- Number of terms in the query (`nterms`): The number of words that compose each query. Mendoza and Baeza-Yates [7] show that a significant proportion of the queries with five or more query terms belongs to the informational category.
- Number of clicks in query sessions (`nclicks`): Average number of selected documents per session calculated over the set of query sessions related to a query (the sessions where the query was formulated). Lee *et al.* [5] show that a significant number of navigational queries concentrates only a few clicks per session.
- Levenshtein distance: distance function calculated among the terms that compose the query and the snippets (the snippet is compounded by the excerpt presented with the query result, the title and the URL of the selected document). Mendoza and Baeza-Yates [7] show that the distance distribution for the informational category has a media of 39.67 calculated over 12,712 pairs queries - snippets. The distribution gotten for non-informational queries (navigational and transactional queries) has a media of 37.6 over 10,540 pairs queries - snippets.

[5] America On-Line Search Engine. Query log files are available for research purposes on *http://www.gregsadetsky.com/aol-data/*

- Number of sessions with less than n clicks over the total of sessions associated to a query (`nCS`): Liu *et al.* [6] introduce the `nCS` feature, that is defined as the number of sessions of a query q that register less than n selections, calculated over the set of sessions where q was formulated. Liu *et al.* calculate the measure for $n = 2$ and $n = 3$. In the extremes of the distribution gotten, this is, for values around 0.95 and 0.05, the informational category exceed the values achieved by non-informational queries.
- Number of clicks before the n-th position of the query ranking (`nRS`): In [6]) Liu *et al.* introduce the `nRS` feature, that is defined as the number of sessions of a query q that register selections only in the top-n results of the answer list of q, calculated over the set of sessions where q was formulated. Liu *et al.* calculate the measure for $n = 5$ and $n = 10$. In the extreme of both distributions, in the class of mark 0.95, non-informational queries exceed informational queries.
- PageRank [8]: Mendoza and Baeza-Yates [7] analyze the hyperlink structure between the pages / sites selected for each category. To do this, they calculate the PageRank[6] measure considering the collection of selected documents in each query category as a subgraph of linked pages / sites. Then, they calcule over each category collection the PageRank measure using a fixed-point algorithm over the matrix of hyperlinks. The authors show that the PageRank values for documents selected in sessions of non-informational queries are higher than the ones selected in informational queries.

In order to illustrate the discriminatory capacity of each of the characteristics considered, we will plot the distribution of the characteristic according to the Broder categorization using our dataset. In the case of the `nCS` and `nRS` characteristics, we have tested versions 2CS, 3CS, 5RS and 10RS, which showed the best results in Liu *et al.* [6]. According to the distributions obtained, the variables that have a greater discriminatory capacity are 2CS and 5RS. They are shown along with the results for `nterms`, `nclicks`, Levenshtein distance and Page Rank, in Figure 1.

As we can observe in Figure 1a) the navigational queries generally have fewer terms than the informational queries. The behavior of this characteristic is not as clear for the transactional class. Figure 1b) shows that some informational queries register more than 9 different sites / pages selected in their sessions. This usually does not occur in the case of navigational or transactional queries. Figure 1c) shows that in general, the Levenshtein distance calculated between query terms and snippets is less in the case of navigational queries than for the other categories. Figure 1d) illustrates that a good amount of informational queries register clicks in pages / sites with low Page Rank, as opposed to transactional or navigational queries. Figure 1e) shows that the characteristic `2CS` is useful for

[6] The PageRank of a page / site is the stationary probability of visiting it in a random walk of the web where the set of states of the random walk process is the set of pages / sites of the web and the transitions between states are one of the following two cases: a) To follow an outgoing link of the page, b) To jump to another page / site selected randomly from the entire web.

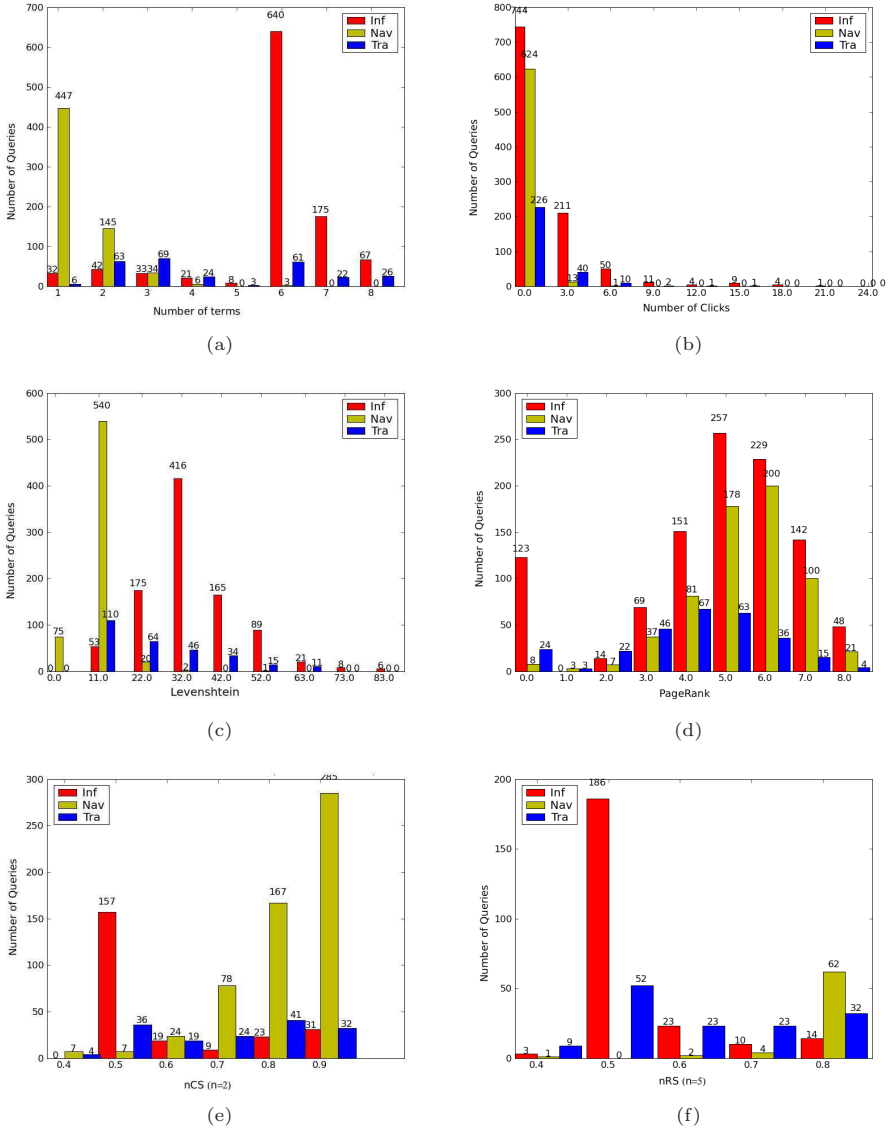

Fig. 1. Feature analysis by category: (a) `nterms`, (b) `nclicks`, (c) Leveshtein distance, (d) PageRank, (e) `2CS` and (f) `5RS`

differentiating between navigational queries. Figure 1f) shows something similar for the `5RS` characteristic, which in this case is useful for differentiating informational queries.

3.3 Building the decision tree

To build the decision tree we use 30% of the data set for evaluation and 70% for training (randomly divided). We use C4.5 to built the decision tree. C4.5 [9], widely used for statistical classification, chooses the feature with the highest normalized information gain to split the training data. Then the algorithm recurses on each training data subset until no more features can be used. The C4.5 implementation corresponds to that provided by the open-source Weka software [7]. The tree was built by Weka in just 0.12 seconds and is shown in Figure 2.

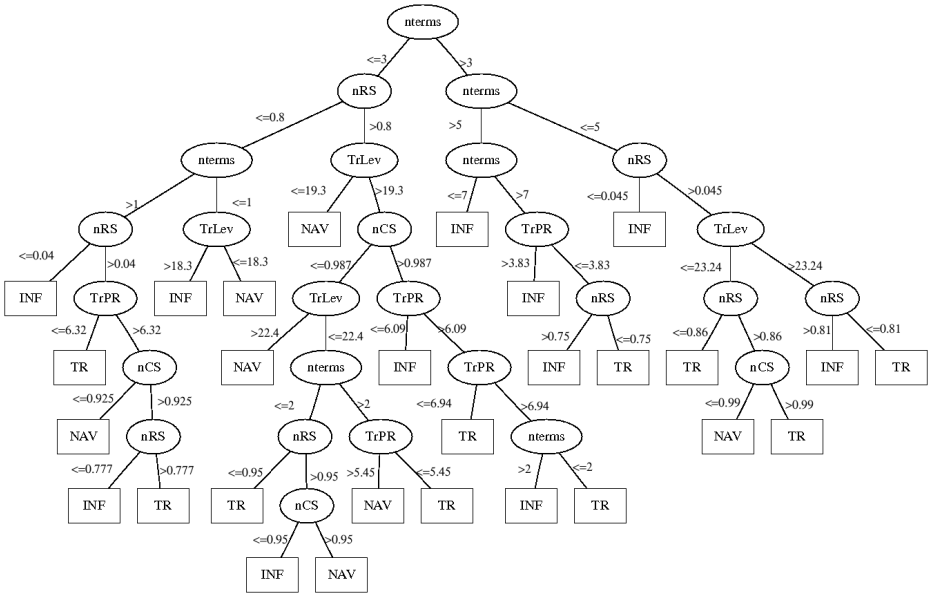

Fig. 2. Decision tree proposed for the problem of query intent identification

We can see that a characteristic relevant to the majority of tree branches is `nterms`, which creates the first partition in the data set. This characteristic can detect informational queries when $3 <$ `nterms`. Combinations with `5RS` can detect the transactional class in this branch. A third case in this branch identifies navigational queries, for which the `2CS` characteristic also needs to be used. When `nterms` ≤ 3, more characteristics are needed to be able to classify. For example, PageRank and Levenshtein distance are used.

[7] Waikato Environment for Knowledge Analysis (Weka), University of Waikato, New Zealand. Available on *http://www.cs.waikato.ac.nz/ml/weka*

4 Evaluation

To consider the costs of evaluation (tradeoff predictive / discriminative) from the comparison of the nominal / predicted class, we consider the four possible cases: true positives (tp), false positives (fp), false negatives (fn) and true negatives (tn). Based on these four cases, the following performance evaluation measurements will be calculated: Precision ($\frac{tp}{tp+fp}$), FP rate ($\frac{fp}{fp+tn}$), TP rate or Recall ($\frac{tp}{tp+fn}$) and F-measure ($\frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}$). In the case of the F-measure, the F_1 version has been considered, which corresponds to the harmonic mean between Precision and Recall, given that it can assess the commitment between both criteria.

The classifier performance will be evaluated by comparing the results from each category analyzed (comparing reference class vs. other classes). The results of this analysis are shown in Table 1.

Comparison	Measures			
	FP Rate	Precision	Recall	F-Measure
(1) Informational - Other Classes	0.182	0.841	0.917	0.878
(2) Navigational - Other Classes	0.066	0.876	0.953	0.913
(3) Transactional - Other Classes	0.032	0.673	0.355	0.465
(4) Weighted Average	0.120	0.826	0.840	0.824

Table 1. Performance evaluation of the proposed decision tree.

As we can see in Table 1, the results are high in both precision and in recall, with an average of 0.824 for the F_1-measure. The lowest performing class is transactional, even though this class obtains a high precision over the portion of queries in this category.

To evaluate the predictive / discriminative tradeoff between unbalanced classes, we use ROC curves (**R**eceiver **O**perating **C**haracteristics), that means to plot TP Rate (benefit) vs. FP Rate (cost). To the extent that the classifier performs well, the area under the curve (AUC) will be greater (maximizing the cost / benefit relationship). The AUC values obtained were as follows: 0.8979, 0.9551 and 0.7357, for the informational, navigational, and transactional categories, respectively. The results obtained show that the predictive / discriminative tradeoff is nearly optimal in the case of navigational and informational categories, and it is acceptable in the transactional case.

5 Conclusions

In this work we have presented a new query classifier according to the Broder taxonomy, based on characteristics and built using decision trees through C4.5

implementation. The resulting tree allows precedence relationships between characteristics to be established. Thus, it can be concluded that all the characteristics considered have been relevant for identifying the Broder categories. Experimental results affirm that the resulting classifier obtains both high precision and recall results, maintaining a balance in the prediction / discrimination relationship, especially for the informational and navigational categories. The case of the transactional category, it has been more challenging to identify characteristics that lead to its clear detection, which has resulted in lower performance in this case. For future work, the exploration of new characteristics and / or classification techniques should be considered so as to improve the results for the transactional category.

Acknowledgments

Dr. Mendoza was partially supported by DIPUV project 52/07 from Universidad de Valparaiso, Chile. Mr. Zamora was supported by a fellowship for scientific initiation of the Graduate School of the UTFSM, Chile.

References

1. R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *Proceedings of SPIRE '06*, pages 98–109, Oct 11th - 13th, 2006, Glasgow, Scotland, Springer LNCS 4209.
2. A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
3. B. Jansen, D. Booth, A. Spink. Determining the informational, navigational and transactional intent of Web queries. *Information Processing and Management*, 44(3):1251–1266, 2008.
4. I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of SIGIR '03*, pages 64–71, Jul 28th - Aug 1st, 2003, Toronto, Canada, ACM.
5. U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proceedings of WWW '05*, pages 391–400, May 10th - 14th, 2005, Chiba, Japan, ACM.
6. Y. Liu, M. Zhang, L. Ru, and S. Ma. Automatic query type identification based on click through information. In *Proceedings of AIRS '06*, pages 593–600, Oct 16th - 18th, 2006, Singapore, Springer LNCS 4182.
7. M. Mendoza, and R. Baeza-Yates. A web search analysis considering the intention behind queries. In *Proceedings of LA-WEB '08*, pages 66–74, Oct 28th - 30th, 2008, Vila Velha, ES, Brazil, IEEE Computer Society Press.
8. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of WWW '98*, pp. 161-172, Brisbane, Australia, 1998, ACM.
9. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publichers, San Mateo, CA., 1993.
10. D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of WWW '04*, pages 13–19, May 17th - 20th, 2004, New York, NY, USA, ACM.