A Simple Model for Classifying Web Queries by User Intent

D. Irazú Hernández[1], Parth Gupta[2], Paolo Rosso[3], and Martha Rocha[1, 4]

[1] Instituto Tecnológico de León, México
[2] Universitat Politècnica de València, Spain
[3] NLE Lab.- ELiRF, Universitat Politècnica de València, Spain
[4] PRHLT, Universitat Politèctica de València, Spain
dirazuhdezfarias@gmail.com
{pgupta,prosso,mrocha}@dsic.upv.es

Abstract. Classifying Web Queries by User Intent aims to identify the type of information need behind the queries. In this paper we use a set of features extracted only from the terms including in the query, without any external or additional information. We automatically extracted the features proposed from two different corpora, then implemented machine learning algorithms to validate the accuracy of the classification, and evaluate the results. We analyze the distribution of the features in the queries per class, present the classification results obtained and draw some conclusions about the feature query distribution.

Keywords: Web query classification, User intent, Query analysis

1 Introduction

Web Search Engines (WSEs) are the most popular tools for access to Internet that people use. According to [5], nearly 70% of the people use a WSE for access to the Web. Identifying the user intent behind a query could help to improve the performance of WSEs, associating the resources available with the user's needs.

Query Classification based on user intent aims to classify queries into categories in relation to the need behind the queries. Jansen and Booth [4], define *user intent* as *the expression of an affective, cognitive, or situational goal in an interaction with a Web Search Engine*. Query Classification based on user intent is different from traditional text classification because of mainly two issues [2]: first, web queries are usually very short; second, many queries are ambiguous and it is common than a query belongs to multiple categories. For example, for the query "*opera theatre tickets*", it is difficult to identify if the user wants to know the website or to buy tickets to attend the event. The most of the efforts have usually involved small quantities of queries manually classified.

Most of the researches on this topic follow the Broder's taxonomy [1], which classifies Web Queries according to their intent into three categories (see Table 1):

Table 1. Taxonomy of User Intent Query Classification

User Intent	Purpose	Example
Navigational	To reach a particular site that user has in mind.	*airport of Chicago*
Informational	To find information assumed to be available on the Web.	*how to apply for passports*
Transactional	To perform further interaction in a site.	*printable maps of nc conties*

The available resources including the query logs, the anchor text, the results returned from WSE together with query text, are usually used to extract features to represent a query. The main contribution of this work is the similar performance obtained with a simple and direct feature extraction method to those of the state-of-the-art [5], [6] and [7].

The remainder of the paper is structured as follows: in Section 2 previous studies are reviewed. Section 3 presents the proposal approach relating the features extracted for *automatic* query representation. In Section 4 experiments and results are described. Finally, conclusions are presented in Section 5.

2 Related Work

There are many approaches to classify user intents. In general they can be divided into several categories [2]. One category tries to augment the queries with extra data, including the search results returned for a certain query, the information from an existing corpus, or an intermediate taxonomy. The second category leverages unlabeled data to help improving the accuracy of supervised learning. The third category expands the training data by automatically labeling some queries in some click-through data via self-training. The anchor text and results from search engines together with query text are used to represent a query. Ganti et al. [3] used tag ratio features for query based on co-occurrence between various types of tags and query terms. Wu et al. [7] presents a dependency relation and word sense features of query text, bigram term and content features for representing a query. Some researches present studies about the features for identification of each type of query. Jansen et al. in [4], [5] and [6], present a methodology developed to classify user intent in terms of the type of content specified by the query and other user expressions, a set of characteristics for each category in Broder's taxonomy and reported three levels of categories in user intent, respectively, the last two are obtained from manually classified queries.

3 Our Approach

Our proposal consists in automatically classifying queries using only the text including in the query. With base mainly in the features described in [1], [5], [6] and [7], we use a set of characteristics for identifying user intent in a query. We can extract other

features from the query text, but this would imply a deeper linguistic analysis that involves consider the query terms in a different way of our proposal. We decided to use as query features only the text that contains because they do not depend on any other source of information rather than the query itself. In previous work it has been shown that the use of text as the only source of information does not allow to obtain such good results as when combined with other sources of information, such as query logs [5]. However, adding this type of information depends on previous queries, and to carry out a process of data collection for the query log. Thus, the use of text as a unique resource for identifying user intent could be done at the time that the user introduces the query to the WSE.

We considering a query q like a set of terms, where each word in the query is independent from each other. For query representation we use a feature vector of query *F(q)*, defined as: $F(q) = \{ EN_q, |q|, T_q, I_q, SW_q \}$, where EN_q are Entity Names, $|q|$ is the query length, T_q and I_q are Transactional, Interrogative terms respectively, and SW_q are stopwords in the query. The EN_q and I_q features are obtained from a POS tagging[1] process, the second feature is obtained by the quantity of terms included in the query, the transactional terms list was defined with words like image, download, buy, sell, and file extensions, this words was included in this list because we found that terms like this are commonly in transactional queries. The last feature was defined from a default stopwords list[2]. We did not make any adaptation of the used tagger, because we only aim at recognizing the features mentioned above, other structure information from the query is not needed.

4 Experiments and Results Discussion

For testing our approach, we used a subset of queries containing in LETOR 4.0[3]. This corpus is a package of datasets for research on LEarning TO Rank, which was released in July 2009. It uses two query datasets from Million Query Track of TREC 2007 (MQ2007) and 2008 (MQ2008). We use the MQ2007 (with 1692 queries) and MQ2008 (with 784 queries). Although we only apply the proposed approach to the above mentioned corpora, it is possible to test our methodology in different data sets where the queries are labeled.

In the next table we present the results of manual classification from both corpora made by inter-annotator agreement. Three annotators were employed: in the first stage, two annotators labeled the corpora, and then we review the classification results. If a query was labeled with a different class, the third annotator assigned a label to the query according to her own judgment. Table 2 shows similar results according to the distribution of user intent queries reported in state-of-the-art.

[1] http://nlp.stanford.edu/index.shtml
[2] http://www.ranks.nl/resources/stopwords.html
[3] http://research.microsoft.com/en-us/um/beijing/projects/letor//

Table 2. Manual Classification of Web Queries by User Intent

	MQ2007	MQ2008
Informational	82%	82%
Navigational	11.5 %	11%
Transactional	6.5%	7%

During the experiments, we removed some features (verbs in the query and domains suffixes terms) that could be extracted from the text according to the state of the art. The features were eliminated because we perform an analysis of discrimination that had each feature and found that some of them were not good discriminators between categories. With this analysis we decreased the dimensionality of our classification process.

In our classification process we use two machine learning algorithms, the Naive Bayes and Support Vector Machine (SVM), in their implementation in Weka[4]. Both algorithms are widely used in text classification. We use the default parameters of SVM in Weka, that implements a sequencial minimal optimization algorithm for training a support vector classifier with a PolyKernel. We performed a 5 cross-validation classification. Tables 3,4 and 5 shows the results of the automatic classification of Web queries. We have used the precision, recall and F-measure to evaluate the performance of algorithms for each user intent category.

Table 3. Automatic Classification of Web Queries by User Intent (Precision)

	Precision					
	Naive Bayes			**SVM**		
	Informational	**Transactional**	**Navigational**	**Informational**	**Transactional**	**Navigational**
MQ2007	0.851	0.734	0.033	0.857	0.734	0
MQ2008	0.929	0.84	0.275	0.867	0.795	0

Table 4. Automatic Classification of Web Queries by User Intent (Recall)

	Recall					
	Naive Bayes			**SVM**		
	Informational	**Transactional**	**Navigational**	**Informational**	**Transactional**	**Navigational**
MQ2007	0.886	0.747	0.088	0.983	0.747	0
MQ2008	0.759	0.810	0.698	0.977	0.810	0

Table 5. Automatic Classification of Web Queries by User Intent (F-measure)

	F-measure					
	Naive Bayes			**SVM**		
	Informational	**Transactional**	**Navigational**	**Informational**	**Transactional**	**Navigational**
MQ2007	0.862	**0.733**	**0.048**	**0.915**	0.733	0
MQ2008	0.834	**0.821**	**0.391**	**0.919**	0.801	0

[4] http://www.cs.waikato.ac.nz/ml/weka/

The algorithms had similar classification performance in both corpora, although SVM obtained better results on the informational category and Navie Bayes is better for others categories. We can see that features extracted from the text allow us to identify transactional queries and have good classification results. However, in the case of navigational queries, results are very low with Naïve Bayes and null with SVM, so we can conclude that the features selected in this work are not enough to represent such queries. We found in previous work that Name Entities allow identifying navigational queries. However, although we use a post-tagging process for the extraction of the features, isolating this characteristic for this kind of queries is not achieved properly. In the feature analysis we review that most of the informational queries contains Name Entities, and this causes that this feature is not a discriminator between these categories. In the classification results, the majority of navigational queries are classified like informational queries.

5 Conclusions

The information behind the queries that a user introduces in a Web Search Engine is very useful for different tasks such as in a Web Search Engine, can improve its performance. Basically the intent of a user query can be classified into three categories: informational, navigational and transactional. There are many approaches for representing the queries: some of them using information obtained from query log, and others add information from different sources. In this work, we use features extracted only from the text contained in the queries. We have tested the query representation using two machine learning algorithms, and we obtained favorable results for classifying informational and transactional queries, but low results for navigational ones. It seems that SVM is more suited for informational queries, and Naïve Bayes for the other two categories.

We reviewed the features distribution per class in the queries and we found that some features work well to separate categories, whereas others have very similar values between them. We can conclude that the use of only the content words in the queries is not enough for classify all the user intents.

Acknowledgements. This research work is the result of the collaboration between the Instituto Tecnológico de León and the Universitat Politècnica de València thanks to the 3-month internship of Irazú Hernández under the grant no. 51188 of CONACYT-Mexico. The work of Parth Gupta and Paolo Rosso was done in the framework of the European Commission WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People, the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-c04-03 (Plan I+D+i) and the VLC/CAMPUS Microcluster on Mutimodal Interaction in Intelligent Systems.

6 References

1. Broder Andrei, A taxonomy of web search, ACM SIGIR, V. 36, Issue 2, pp. 3-10, 2002.
2. Cao Huanhuan, Hao Hu Derek, Shen Dou, Jiang Daxin, Sun Jian-Tao, Chen Enhong, Yang Qiang, Context-Aware Query Classification, The 32nd Annual ACM SIGIR Conference, pp. 3-10, 2009.
3. Ganti Venkatesh, König Arnd Chistian, Li Xiao, Precomputing Search Features for Fast an Accurate Query Classification, Proceedings of the third ACM International Conference on Web Search and Data Mining ACM, pp. 61-70, 2010.
4. Jansen Bernard J., Booth Danielle, Classifying Web Queries by Topic and User Intent, Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, pp. 4285-4290, 2010.
5. Jansen Bernard J., Booth Danielle L., Spink Amanda, Determining the informational, navigational, and transactional intent of Web queries, Journal Information Processing and Management: an International Journal archive V. 44 Issue 3, pp. 1251-1266, 2008.
6. Jansen Bernard J., Booth Danielle L., Spink Amanda, Determining the User Intent of Web Search Engine Queries, Proceedings of the 16th international conference on World Wide Web ACM, pp. 1149-1150, 2007.
7. Wu Dayong, Zhang Yu, Zhao Shiqi, Liu Ting, Identification of Web Query Intent Based on Query Text and Web Knowledge, Pervasive Computing Signal Processing and Applications (PCSPA), First International Conference, pp. 128-131, 2010.