ANALYSIS OF THE SYNTACTICAL STRUCTURE OF WEB QUERIES

ALAA MOHASSEB[1], MOHAMED BADER-El-DEN[1], MIHAELA COCEA[1]

[1]School of Computing, University of Portsmouth
E-MAIL: alaa.mohasseb@port.ac.uk, mohamed.bader@port.ac.uk, mihaela.cocea@port.ac.uk

Abstract:

The analysis of web queries is an important task in the enhancement of the identification of users' search intent. This task has been addressed by various studies; many of these studies automatically classified queries based on query characteristics, users' behaviour and words. In this paper, we present an analysis of web queries based on their syntactical structure. We identify different patterns and use machine learning algorithms to classify them according to Broder's categories of user intent, i.e. informational, navigational and transactional. Experimental results show that our approach has a good classification performance, especially for informational and navigational queries.

Keywords:

Query Analysis, Query Classification, Information Retrieval, user intent, Natural language processing

1 Introduction

Search engines are the most popular Information Retrieval applications. Despite that search engines try to improve the user experience and the technology used in finding relevant results, many difficulties are still faced because of the continuous increase in the amount of web content.

One major task in identifying the intent of a user's query is the classification of the query type. There are several taxonomies of web queries [1, 2, 3, 4, 5, 6, 7], of which Broders taxonomy [2] is one of the most commonly used. It includes three main types: informational, navigational and transactional queries. The analysis of web queries has been addressed in many studies, many of which classified queries by using: (a) the characteristics of each query type [2, 8, 9, 10], (b) users' behaviour by analyzing the query logs [11, 4, 7, 12, 13] and (c) click through data [14, 6, 15, 16]. Furthermore, research such as [17] and [18] analyzed the linguistic structure of web queries by applying techniques from natural language processing, such as part of speech tagging. Web query could be structurally complex [17], leading to the fact that two queries with overlapping sets of terms may reflect two totally different intents. To distinguish between these, users' behaviour or user clicks were used; however, these alone could be misleading in identifying the intent of a query [13].

In this paper we present a syntax-based approach for analysing the structure of web queries. Several patterns of different syntactic structures are identified and machine learning is used for the automatic classification of these patterns according to Broder's taxonomy of user intent (i.e. informational, navigational and transactional).

The rest of the paper is organized as follows: Section 2 provides an overview of related work, including user intent taxonomies and approaches for query analysis and classification. Section 3 presents our proposed approach for query analysis and automatic classification. The evaluation experiments and results are described and discussed in Section 4, and Section 5 concludes the paper and outlines directions for future work.

2 Related Work

In this section we review related work on: (a) query taxonomies for user intent; (b) methods for web queries' analysis and classification.

2.1 Web Queries Taxonomies for User Intent

There are many different proposed taxonomies of web queries. In our research, the web queries taxonomy proposed by [2] is used due to its popularity in previous research. Broder's categorized the queries according to users intent into three categories: informational, navigational and transactional.

Topics related to informational queries are broad and general, or quite specific and there are no particular web pages containing all the information needed; users have to acquire this information from multiple web pages. The objective of this type of search is to answer a question or to find information in order to learn how to do something. On the other hand, the purpose of transactional queries may be to acquire information about something or to find a site and further interaction may be

required, such as downloading software, buying a certain product online. Finally, the objective of navigational queries is to reach a particular site and usually this type of queries have just one right result.

2.2 Methods for Web Queries Analysis and Classification

Different analysis methods have been used to classify queries and to identify users' search intent by using. [7] analysed the logs of a commercial web search engine and studied the web search queries for their diversification requirements. Similarly, [12] used query logs to identify a list of categories which can describe a given query. [13] used search logs for the purpose of enriching a query by mining the previous documents clicked by users and the relevant follow up queries in a session; a text classifier was used to map the documents and the queries into predefined categories.

The query analysis by [19] was done by using two types of features: past user click behavior and Anchor-link distribution, while, [20], [10] and [21] used a variety of query features to automatically classify the user intent behind web queries. Furthermore, [8] also classified user intent based of the characteristics of the queries. [22] research involved the analysis of the semantic structure of noun phrase queries. Furthermore, [17] examined the structure of web queries by applying techniques from natural language understanding. Finally, [18] analysis of queries was based on the syntax of part of speech tag sequences. Their results showed that query part-of-speech tagging can be used to create significant features for improving the relevance of web search results and may assist with query reformulation.

3 Proposed Approach

In this section, we start by describing the data used in our analysis, as well as the syntactic categories used in the analysis of the queries. We also describe how the syntactic categories were used to create syntactic patterns for each query type (i.e. informational, navigational and transactional).

3.1 Queries Syntactical Structure

Most of the queries submitted to search engines might have more than one meaning, therefore using only the terms to identify search intents is not enough. To address this problem, we explore the syntactic structure of queries.

Two different queries may have similar terms but with different structures, each having a different meaning, which may lead to different intents. For example, both queries *George Orwell books order* and *order George Orwell books* have similar terms and by just looking at them, one might assume that for both the intent is to buy books, i.e. transactional intent. According to the characteristics of the informational, navigational and transactional intents from [2], the first query is informational (i.e. find information on George Orwell books), while the second query is transactional (i.e. buy George Orwell books). We illustrate below how the syntactical structure of the queries can reflect these different intents.

A phrase, defined as a group of words that function as a single part of speech, can be a Verb phrase, Noun phrase, Determiner phrase, Adjective phrase, Adverb phrase or Prepositional phrase. Different classes of phrases contain different word classes. A word class or part of speech is a collection of words that can have subclasses; the seven major word classes are Verb, Noun, Determiner, Adjective, Adverb, Preposition and Conjunction. Word order inside a phrase is one of the major structural ways in which the queries can differ from each other. The position of a word depends on its word class, which means that each query could formulate a unique pattern.

At word level, *"George Orwell books order"* consists of *Nouns*, while *"order George Orwell books"* consists of a *Verb* and *Nouns*. At phrase level, *"George Orwell books order"* consists of *Noun Phrases*, while *"order George Orwell books"* consists of a *Verb Phrase* and a *Noun Phrase*. This different syntactical structure of the two queries leads to different syntactical patterns, which result in different meaning, intent and search results.

The following categories/word classes have been used, Verb (V), Noun (N), Determiner (D), Adjective (Adj), Adverb (Adv), Preposition (P) and Conjunction (Conj). In addition, question words (QW): how, who, when, where, what and which, were also used. Furthermore, we also added two other classes: Domain Suffixes (DS) and Prefixes (DP). Also, some word classes can have subclasses. For example, Nouns consists of subclasses, such as Common Nouns (CN), Proper Nouns (PN), Pronouns (Pron) and Numeral Nouns (N); Verbs can be of several types, such as Action Verbs (AV), linking Verbs (LV) and Auxiliary Verbs (AuxV).

4 Analysis of Query Types

We have looked at the characteristics of the three different types of queries, i.e. informational, navigational and transactional, from the point of view of the different word classes and types of phrases reflected in these queries. The analysis of web search queries syntactical structure were done through exam-

ining 50,000 randomly selected queries from the AOL 2006 data-set[1] [23] and the TREC 2009 Million Query Track data-set[2] [24]. Details for each query type are given below.

4.1 Informational Query

One of the main feature that identifies the structure of informational queries is Phrases such as Noun phrase (NP), Verb phrase (VP), and Prepositional phrase (PP). For example *"location of apple stores in London"*. The most used word class in this query type is Nouns, such as Common Nouns, e.g. *"county", "company"* and *"place"*, and Proper Nouns, such as *"Spain"*, *"Eiffel Tower"* and *"The Beatles"*. Question words are also used; for example *"Why recycling is important?*; informational query is the only type of queries that contain Question words. Moreover, queries in such search type could be short, medium or long in length, and they could contain one word or more than five words [8]. Furthermore, informational queries mostly formulate a complete sentence such as *"where can i buy vegan products in the UK?"*. However, in many cases informational queries could be short in length [8], such as *Dinner ideas"*. Two examples of informational search syntactical structures are shown in Figure 1.

4.2 Navigational Query

The structure of the query is the main feature that distinguishes navigational queries. This type of queries normally have a fixed syntactical structure which is the Noun Phrase (NP). Also, in some cases the query contains a web-link or part of a web-link. Furthermore, queries in this search type are mainly short, consisting of one or two words only [8]. Moreover, the only sub-class that could be found in this type of query is Proper Nouns since the query could contain just one word typically containing an organization, business, company or university name, such as *"Microsoft"*. In addition, the structure of the query consist of domain suffixes and prefixes such as *"https://www.google.co.uk"* or *"amazon.com"*, as shown in Figure 2.

4.3 Transactional Query

The syntactic structure of transactional queries consists mostly of Verb Phrases (VP) and Adjective Phrases (AP) for example *"buy cheap phones"*. Also, Noun Phrases (NP) could be in the structure of some queries – for example *"Phil Collins lyrics"*; however, some word classes are not used such as Question words, Pronouns, and Auxiliary verbs. Moreover, most queries in transactional searching consist of Action Verbs (AV) such as *"order"*, *"buy"*, *"purchase"*, and *download"*. Furthermore, Adjectives are one of the word classes being used frequently in transactional queries, such as *"Free"* and *"online"*. In addition, queries in this search type could be short or medium [8], they could contain one word or up to five words – for example *"cupcakes recipes"* and *"online pdf to word converter"*. Figure 3 shows an example of a query structure for a transactional query.

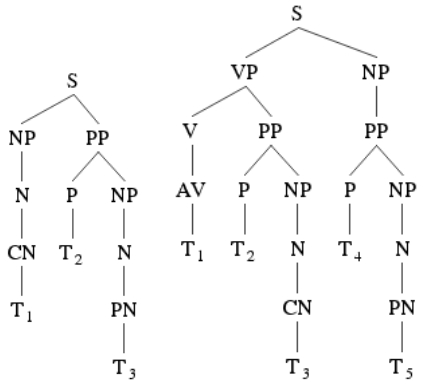

FIGURE 1. Examples of informational query structure using syntax tree representation, in which each sentence consists of a syntax structure of phrases (NP, PP, VP), word classes (N, V, P) and word sub-classes (PN, CN, AV); a sentence could have more than one of each.

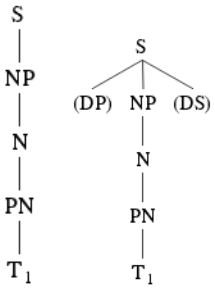

FIGURE 2. Examples of navigational query structure using syntax tree representation; the two patterns displayed cover the most common queries in the Navigational search. The sentences could consist of domain suffixes or prefixes (DS, DP), or have a syntactic structure of phrases (NP), word classes (N) and words sub-classes (PN).

[1] http://www.researchpipeline.com/mediawiki/index.php?title=AOL_Search_Query_Logs

[2] http://trec.nist.gov/data/million.query09.html

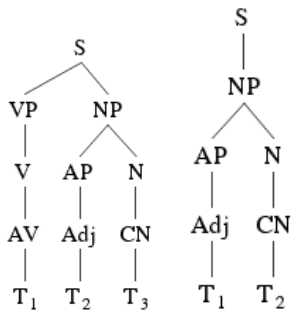

FIGURE 3. Example of a transactional query structure using syntax tree representation, in which each sentence consists of a syntactic structure of phrases (NP, AP, VP), word classes (N, V, Adj) and word sub-classes (CN, AV); a sentence could have more than one of each.

4.4 Analysis overview

Based on the analysis above, an overview of the syntactical structure characteristics of the informational, navigational and transactional search type queries is presented in Tables 1,2, 3 and 4.

Table 1 outlines the difference between the three types of queries from the point of view of word classes and Table 2 shows the types of phrases present in the three different query types. Both tables show that the navigational queries are clearly different from the other two, while the informational and transactional queries have a large similarity, indicating the difficulty in distinguishing them.

Table 3 outlines the difference between the three types of queries bases on different types of verbs. Navigational queries do not typically contain verbs, while the informational ones do. Moreover, the transactional queries tend to contain a particular type of verb, i.e. Action Verb (AV), but not the others, thus indicating that this particular verb class plays an important role in the identification of transactional queries.

Table 4 outlines the different types of nouns present in the three query types. Transactional queries tend not to include pronouns, while the navigational queries typically do not include Common Nouns and Numeral Nouns.

5 Experiments and Results

To investigate the ability of machine learning classifiers to distinguish between informational, navigational and transactional queries, we used 20,000 queries were randomly selected from AOL 2006 data-set and TREC 2009 Million Query Track data-set. The selected queries were different from those used in the identification of the query syntactical patterns for the three types of user intent (informational, navigational and transactional). Two machine learning algorithms, i.e. *Random Forest* and *Naive Bayes*, were used for query classification due to their popularity in text classification

To assess the performance of the machine learning classifiers, the Weka[3] software [25] was used. The experiments were set up using the typical 10-fold cross validation, i.e. the dataset is split into 10 folds, and each fold in used, in turn, for testing, while the other 9 are used for training. The output of the training process is a model, which is then used to classify the queries in the test fold. The labels produced by the model are matched to the true labels and typical performance indicators, such as accuracy, precision, recall, and F-score, are calculated. The results are presented in the next subsection.

5.1 Results

Tables 5 and 6 display the precision, recall and F-score for the Random Forrest and Naive Bayes classifiers, respectively. In addition, the accuracy for the Random Forest model is 86.14%, while for the Naive Bayes it is 80.12%.

Random Forest incorrectly classified 13.86% of the queries: (a) 8.24% of the informational queries were classified as transactional and 0.84% as navigational; (b) 4.44% of the transactional queries were classified as informational and 0.35% as navigational; (c) navigational queries were 100% correctly classified. The Naive Bayes classifier incorrectly classified 19.88% of the queries: (a) 13.47% of the informational queries are classified as transactional and 0.57% as navigational; (b) 4.95% of the transactional queries are classified as informational and 0.13% as navigational; (c) 0.76% of the navigational queries are classified as informational.

Comparing the effectiveness of the classifiers, the Random Forest classifier has the highest precision, recall and F-score for the informational and transactional queries. Regarding the navigational queries, Naive Bayes has the highest precision, while Random Forest has the highest recall and F-score.

The results indicate that queries can be automatically classified into the three different types of user intent with a good level of performance. Two of the user intents, i.e. informational and transactional, are easier to identify than the transactional queries. In addition, this experiment validates the ability of our approach to automatically identify and classify query types using the syntactical structure of the queries.

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

TABLE 1. Analysis of Word classes (Part of the Speech)

Queries	Structure Length			Word classes							
	S	M	L	N	V	D	Adj	Adv	P	Conj	QW
Informational Query	√	√	√	√	√	√	√	√	√	√	√
Navigational Query	√	-	-	√	-	-	-	-	-	-	-
Transactional Query	√	√	-	√	√	√	√	√	√	√	-

TABLE 2. Analysis of Phrases

Queries	NP	VP	PP	AdvP	AdjP
Informational Query	√	√	√	√	√
Navigational Query	√	-	-	-	-
Transactional Query	√	√	√	√	√

TABLE 3. Breakdown Analysis of the Verb Class

Queries	AV	AuxV	LV
Informational Query	√	√	√
Navigational Query	-	-	-
Transactional Query	√	-	-

TABLE 4. Breakdown Analysis of the Noun Class

Queries	CN	PN	Pron	NN
Informational Query	√	√	√	√
Navigational Query	-	√	√	-
Transactional Query	√	√	-	√

TABLE 5. Random Forest results

Query Search Type	Precision	Recall	F-score
Informational Query	0.936	0.878	0.906
Navigational Query	0.867	1.000	0.929
Transactional Query	0.613	0.731	0.667
Overall	0.805	0.870	0.834

6 Conclusion

In this paper, an analysis of web search queries was provided by identifying the syntactical structure of each type of search query, i.e. informational, transactional and navigational. Furthermore, the impact of using the queries' syntactical structure on the automatic query classification performance was tested and showed that our approach outperformed most of the existing approaches.

In future work, we will examine and analyze more queries from different search engines to extend the ability of our system to identify more queries. We will also extend the analysis of the syntactical patterns to include domain-specific information, and investigate the influence of this additional type of information of the query classification performance.

References

[1] J. B. Morrison, P. Pirolli, and S. K. Card, "A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions," in *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 2001, pp. 163–164.

[2] A. Broder, "A taxonomy of web search," *ACM Sigir forum*, vol. 36, no. 2, pp. 3–10, 2002.

[3] M. Kellar, C. Watters, and M. Shepherd, "A goal-based classification of web information tasks," *Proceedings of the American Society for Information Science and Technology*, vol. 43, no. 1, pp. 1–22, 2006.

[4] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro, "The intention behind web queries," in *International Symposium on String Processing and Information Retrieval*. Springer, 2006, pp. 98–109.

[5] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo, "Classifying and characterizing query intent," in *European Conference on Information Retrieval*. Springer, 2009, pp. 578–586.

[6] D. Lewandowski, J. Drechsler, and S. Mach, "Deriving query intents from web search engine queries," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 9, pp. 1773–1788, 2012.

[7] S. Bhatia, C. Brunk, and P. Mitra, "Analysis and automatic classification of web search queries for diversification requirements," *Proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–10, 2012.

[8] B. J. Jansen, D. L. Booth, and A. Spink, "Determining

TABLE 6. Naive Bayes results

Query Search Type	Precision	Recall	F-score
Informational Query	0.914	0.811	0.859
Navigational Query	0.909	0.901	0.905
Transactional Query	0.486	0.715	0.579
Overall	0.770	0.809	0.781

the informational, navigational, and transactional intent of web queries," *Information Processing & Management*, vol. 44, no. 3, pp. 1251–1266, 2008.

[9] D. Wu, Y. Zhang, S. Zhao, and T. Liu, "Identification of web query intent based on query text and web knowledge," in *Pervasive Computing Signal Processing and Applications (PCSPA), 2010 First International Conference on*. IEEE, 2010, pp. 128–131.

[10] L. Calderón-Benavides, C. González-Caro, and R. Baeza-Yates, "Towards a deeper understanding of the users query intent," in *SIGIR 2010 Workshop on Query Representation and Understanding*, 2010, pp. 21–24.

[11] D. E. Rose and D. Levinson, "Understanding user goals in web search," in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 13–19.

[12] C. Højgaard, J. Sejr, and Y.-G. Cheong, "Query categorization from web search logs using machine learning algorithms," *International Journal of Database Theory and Application*, vol. 9, no. 9, pp. 139–148, 2016.

[13] R. Song, Z. Dou, H.-W. Hon, and Y. Yu, "Learning query ambiguity models by using search logs," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 728–738, 2010.

[14] S. M. Beitzel, E. C. Jensen, O. Frieder, D. Grossman, D. D. Lewis, A. Chowdhury, and A. Kolcz, "Automatic web query classification using labeled and unlabeled training data," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 581–582.

[15] M. Mendoza and J. Zamora, "Identifying the intent of a user query using support vector machines," in *International Symposium on String Processing and Information Retrieval*. Springer, 2009, pp. 131–142.

[16] Y. Liu, M. Zhang, L. Ru, and S. Ma, "Automatic query type identification based on click through information," in *Asia Information Retrieval Symposium*. Springer, 2006, pp. 593–600.

[17] R. Saha Roy, "Analyzing linguistic structure of web search queries," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 395–400.

[18] C. Barr, R. Jones, and M. Regelson, "The linguistic structure of english web-search queries," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 1021–1030.

[19] U. Lee, Z. Liu, and J. Cho, "Automatic identification of user goals in web search," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 391–400.

[20] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink, "Classifying the user intent of web queries using k-means clustering," *Internet Research*, vol. 20, no. 5, pp. 563–581, 2010.

[21] A. Figueroa, "Exploring effective features for recognizing the user intent behind web queries," *Computers in Industry*, vol. 68, pp. 162–169, 2015.

[22] X. Li, "Understanding the semantic structure of noun phrase queries," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 1337–1345.

[23] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search." in *InfoScale*, vol. 152, 2006, p. 1.

[24] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas, "Million query track 2009 overview." in *TREC*, 2009.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.