Computers in Industry 68 (2015) 162–169

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Exploring effective features for recognizing the user intent behind web queries

Alejandro Figueroa [a,b,*]

[a] Yahoo! Research Latin America, Blanco Encalada 2120, Santiago, Chile
[b] Escuela de Ingeniería Informática,Universidad Diego Portales, Santiago, Chile

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 14 January 2015
Accepted 15 January 2015
Available online 7 February 2015

Keywords:
Search query understanding
Query classification
Query analysis
User intent
User experience
Feature analysis

ABSTRACT

Automatically identifying the user intent behind web queries has started to catch the attention of the research community, since it allows search engines to enhance user experience by adapting results to that goal. It is broadly agreed that there are three archetypal intentions behind search queries: navigational, resource/transactional and informational.

Thus, as a natural consequence, this task has been interpreted as a multi-class classification problem. At large, recent works have focused on comparing several machine learning methods built with words as features. Conversely, this paper examines the influence of assorted properties on three classification approaches. In particular, it focuses its attention on the contribution of linguistic-based attributes. However, most of natural language processing tools are designed for documents, not web queries. Therefore, as a means of bridging this linguistic gap, we benefited from caseless models, which are trained with traditionally labeled data, but all terms are converted to lowercase before their generation.

Overall, tested attributes proved to be effective by improving on word-based classifiers by up to 8.347% (accuracy), and outperforming a baseline by up to 6.17%. Most notably, linguistic-oriented features, from caseless models, are shown to be instrumental in narrowing the linguistic gap between queries and documents.

1. Introduction

Each day, most popular search engines answer millions of queries, submitted from all over the world by people from all walk of life. Broadly speaking, users express their requests by entering short sequences of query terms, i.e. normally 2–3 keywords [1,2]. Search engines, and information retrieval engines in general, are then challenged to adequately understand these short texts in order to produce better results and enhance user experience.

Discriminating the user intention is an initial step to draw valid query interpretations. In a broad sense, this intent varies from fulfilling information needs to utilize search engines as navigational tools to reach specific web sites they want to access. In addition, search engines can be used to access resources including maps, lyrics, and books. Hence, the automatic detection of user intentions aids in enhancing user experience by returning more useful results to searchers and tailoring these results to their specific needs.

On the one hand, the intent of some highly frequent queries (e.g. "*wikipedia*" and "*yahoo*") can be easily identified by analyzing click patterns across search logs. In a similar manner, the user intention behind queries that observe a limited set of patterns, including *define term* and *person name pics*, can also be readily recognized. It is hard, on the other hand, to determine the intention of the large portion of new queries or wordings by using this kind of heuristic approach, especially if there is no click history available.

Broadly speaking, recent research has concentrated on testing several multi-class classification algorithms [3]. But, by and large, these approaches have been built on top of lexical features, leaving assorted attributes largely unexplored. In a special manner, linguistic-based features have not been studied, since most NLP tools are designed for documents, not for web queries.

The linguistics of documents and web queries differ noticeably. Simply put, search queries provide little context, they bear two or three words on average, therefore they are relatively short with respect to documents [1,4]. Another key difference is that capital letters can signal proper nouns across documents; whereas queries consist predominantly of lowercased terms. Since entities are key to identify user intent, researchers are bridging this gap by devising named entity recognizers specialized in web queries [5–7].

* Correspondence to: Yahoo! Research Latin America, Blanco Encalada 2120, of. 400, Santiago, Chile. Tel.: +56 2 29784632.
E-mail addresses: afiguero@yahoo-inc.com, alejandro.figueroa@mail.udp.cl

http://dx.doi.org/10.1016/j.compind.2015.01.005
0166-3615/

The innovative aspect of our work is to study the effectiveness of assorted fine-grained properties in detecting the user intent behind web queries. More exactly, their impact on three multi-class learners: Naïve Bayes (Bayes), SVM multi-class (SVM) and maximum entropy (MaxEnt). In brief, the features studied in this work comes from:

1. We reduce the linguistic gap by exploiting caseless models, that is to say models trained with conventionally annotated corpora (e.g., a treebank), but before training and creating these models, terms are lowercased. These models include POS tagging, name entity recognition and dependency parsing.
2. We profit from a named entity recognizer devised for queries (NERQ), and from tools particularly useful for classifying short texts, e.g., explicit semantic analysis.

In a nutshell, results show that caseless models are a cost-efficient solution to obtain effective linguistic-based properties from queries, but still yet, purpose-built tools are preferable.

This paper is organized as follows: Section 2 discusses the related work, Section 3 dissects our corpus acquisition strategy, next Section 4 describes our feature set, Section 5 deals at length with our experimental settings and results, and lastly Section 6 draws some conclusions.

2. Related work

In [8], search queries were manually categorized in congruence with a taxonomy, which at its first level is separated into three canonical branches. Each of these three branches denotes a goal that users have in mind when searching: navigational, information, and resource.

Initially, [9] automatically classified web queries into these three user intents. They randomly selected and manually annotated queries distilled from seven search engine logs. They exploited two fertile sources of discriminative features: keywords and information taken from the results pages viewed by the user. They discovered that navigational queries are short in length and typified by the user viewing the first result page. In addition, they found out that this kind of query is characterized by organization and people names, domain suffixes as well. Conversely, resource queries are more likely to embody keywords like lyrics, recipes, and images. In contrast, informational queries are longer, many times formulated with question words, and they are likely to resemble natural language text. By manually examining 400 queries, they found out that the intent of about 75% of these queries is unambiguous for the human reader.

Along the same lines, [10] categorized web queries into these three canonical classes using k-means clustering in conjunction with a variety of query traits. Each of the 4,056,375 records in their search log comprised the type of content collection the user is searching, user identification, cookie, time of day, and query terms. They enriched each record with query length, a number representing the search engine results page viewed during a given interaction, and the number of times a user changed the query during a session. The assignment of terms as informational, navigational, or transactional was based on [9,11].

In the same vein, [12] manually tagged 20,000 search queries in agreement with their user intent. Then, new web queries were automatically classified using an exact match. Since this approach mainly matches frequent queries, it was combined with an n-gram language model. Later, [3] profited from an SVM and a Bayes classifier trained with query terms features. Their corpora encompassed ca. 2500 search queries. They found out that SVM obtained better results on the informational category while Bayes performed better on the other two types. For their largest dataset, their models achieved a precision of 0.857, 0.734 and 0.033 when targeting the informational, resource and navigational categories, respectively. They discovered that word features are good at identifying resource queries, but results are poor for navigational queries. They conjectured that recognizing named entities is necessary for drawing that distinction.

Incidentally, recent studies have shown that the linguistics behind search queries is different from that of text documents [2]. For instance, about 70% of query terms are nouns and proper nouns, adjectives about 7% and URLs 6%. This is in sharp contrast to documents, where almost each sentence contains at least one verb. The use of capital letters is also different: while in documents they are utilized chiefly for signaling proper nouns, in search queries the use of uppercase is inconsistent. Since these dissonances pose a great challenge to traditional NLP tools, researchers have started to design purpose-built tools for tackling web queries, i.e., named entity recognizers [1,4,13–16].

As a means of improving performance, some studies have tried to exploit the context provided by user sessions [17,18], search-result snippets and click-through data [19]. However, automatically identifying user sessions is not an easy task [20], and these can involve users trying to achieve several goals. Note that session-based approaches need to detect the user intent taking into account only one query, when the session starts from scratch, falling back on some generic model.

In contrast, our work shows the beneficial impact of assorted attributes on three multi-class classifiers (i.e., SVM, MaxEnt and Bayes). More precisely, these features were derived from: (a) NLP tools constructed specifically for coping with queries, i.e., NERQ; (b) models trained on caseless corpora, this way we reduce the linguistic gap between search queries and documents; and (c) two semantic query expansion strategies. Results show their effectiveness in improving the detection of the user intent behind web queries.

3. Corpus acquisition

We took advantage of the AOL web query collection,[1] gathered during March and May 2006 [21]. This consists of about 21 million of search queries prompted by approximately 650,000 users. Each instance has an user id, timestamp, search query string, the rank and the URL of the clicked results. More specifically, this collection comprises ca. 10 millions unique lowercased queries, of which 4,811,638 elements are linked with at least one clicked URL. In our study, we capitalize only on these 4.8 millions queries as these clicked URLs facilitate the annotation process. In detail, we performed two sorts of annotations: automatic and manual.

3.1. Automatic annotations

Previous works noticed that the intent of some queries is easy to detect by means of discriminative keywords [9]. Therefore, we also capitalized on discriminative terms. In the event of navigational queries, we sought for words including "*http*" and "*www*". We also checked if the prompted query without whitespaces was contained in one of the clicked URLs, or if it ended with "*.com*". Regarding resource queries, similar to previous works, we examined if the last token was "*pics*", "*lyrics*", "*picture*", "*photos*", "*image*", "*images*", "*recipe*", "*recipes*", "*pictures*", "*map*", "*maps*", "*video*", "*videos*", "*software*", "*cheat*", "*cheats*", "*download*", "*downloads*", "*guide*", "*nude*", "*porn*", "*pdf*", "*7z*", "*deb*", "*gz*", "*rar*", "*tar.gz*", "*cpp*", "*dll*", "*ttf*", "*xml*", "*exe*", "*xlsx*", "*bmp*", "*gif*", "*jpg*", "*png*", "*ps*", "*mpg*", "*wmv*", "*bmp*", "*m4v*", "*mov*", "*mp4*", "*asf*", "*avi*", "*flv*", "*wav*",

[1] gregsadetsky.com/aol-data.