Feature Article:

"Phrases and Web Retrieval"

By Gilad Mishne & Maarten de Rijke

It is generally accepted that stating information needs in a more focused way leads to more precise results. For example, for a tourist arriving with her dog to the Netherlands, the query pet-friendly hotels in Amsterdam will probably yield more accurate results than a search for hotels in Amsterdam. This is why intuition leads us to believe that the usage of phrases when expressing information needs will improve the results of the search: while a search for white house pictures may return both pictures of the residence of the U.S. president and pictures of various houses colored in white, the more focused search for "white house" pictures will narrow the results to the presidential residency by guaranteeing that the words "white" and "house" appear consecutively, in that order, in the retrieved information. Now, it can be argued that this improved performance comes at the cost of missing some truly relevant results (in technical terms, increases precision at the price of decreased recall); however, when dealing with large quantities of information such as those existing on the World Wide Web, this is not an important factor. Studies show that users are concerned mainly with the accuracy of the first 10 or 5 results displayed: they wish to get "some relevant data" ranked highly, and not necessarily "all relevant data."

Following this intuition, the use of phrases has been researched extensively in the last 30 years in the field of Information Retrieval. Sadly, the general conclusion reached was that given a good ranking formula, the use of phrases has no substantial effect on performance. This has been validated a large number of times, using various experimental settings and measurement metrics. However, an important characteristic of all these experimental environments was that the collection of documents on which the search was performed was basically plain-text documents, usually newswire feeds of a number of years.

In this article, we revisit the usage of phrases (and additional, less restrictive multi-word expression units) in retrieval. This time, however, we focus on web retrieval: searching in HTML documents. The questions we set ourselves are:

- Is the use of phrases beneficial in the web setting?
- If so, why is it different from plain-text documents?
- Is there a simple, robust way of automatically using phrases to improve search?

Web documents and phrases

Why do we hypothesize that web documents may exhibit a different behaviour with phrase searches than plain-text documents? To answer this, we point out a prominent feature of HTML pages, often used in the web retrieval setting. HTML was originally devised to describe the physical layout of a page, marking entities such as page title, emphasized text, font size and so on. This markup has been successfully used in numerous ways to improve search performance from web pages, for example by assigning a higher weight to headline text than to paragraph text; this technique is sometimes referred to as multiple document representation, meaning that the document is separated to different "representations" (or fields) — the title, the headline text, the anchor text (text present in all links in pages pointing to a certain page), and so on. Different fields are then searched separately, and results are combined to form a single ranked document list, assigning more importance to representations such as headline text.

Examining these different fields, we see that unlike plain-text documents, some of them are highly rich in phrases: such are the page title, the anchor text, and even the URL text. This is a direct result of the purpose of these fields: the page title is a short, usually human generated summary of the topic of a page, and in many cases is or contains phrases. Similarly, the anchor text and URL text are very short descriptions of the page, meant to summarize its content to a single soundbyte.

Additional short, descriptive fields are the META "keywords" and "description" sometimes associated with web pages, and containing a short list of important words and phrases relevant to the page.

In addition to the difference in the searched documents, in the case of web searches the queries themselves are also rich in phrases. The informational web queries used in evaluations such as TREC contain a very high percentage of phrases (almost 80% of the queries with more than one word are or contain phrases); in the case of web query logs from actual commercial search engines, phrases are also found in the majority of the queries (although percentages are slightly lower).

The combination of the abundance of phrases in certain fields of web documents, and their frequent usage in queries on the web leads us to hypothesize that the use of phrases for searching the web will result in a higher gain to retrieval performance than in the case of plain-text documents that has been explored extensively in the literature.

Detecting phrases and expanding queries

There are three general approaches to phrase recognition in text: *syntactical*, *statistical*, and *lexical*. Syntactical approaches rely on linguistic analysis of the text, using features such as part-of-speech tags or dependency relations to identify words which constitute a phrase. Statistical approaches rely on *collocation* information: phrases are identified by looking at the entire corpus and examining which words tend to appear together. Finally, lexical approaches are based on pre-constructed lists of typical phrases in a language. Naturally, hybrid methods exist which combine ideas from the three techniques.

Given the average length of web queries, and the fact that they are mostly ungrammatical, using the syntactical approach is unlikely to yield good results. Statistical approaches generally provide good results, but need to be adjusted per corpus separately: words that tend to appear together in one collection might not do so in another. Finally, lexical approaches require obtaining large amounts of phrases in advance and are not scalable. We therefore take a simpler, perhaps naïve approach to phrase detection: we consider *every* combination of consecutive words from the query (of any length) as a phrase. For example, given the user-supplied query *well water contamination*, we identify all of the following as phrases: "well water", "water contamination," and "well water contamination." All phrases detected this way are then added to the original query as phrase terms, creating a long query containing the original terms as well as the more focused phrases. With modern retrieval ranking formulas, this results in rewarding documents which contain as many phrases as possible from the query (and still matching documents which contain the separate words, but not the phrases).

The reasoning behind this seemingly shallow approach is as follows:

- Empirically, this approach captures the cast majority of phrases in the queries.
- While this method also makes mistakes, they are unlikely to affect the search performance. For example, assume that we are given the query *automobile emissions vehicle pollution*; while we successfully capture the phrases "automobile emissions" and "vehicle pollution," we incorrectly identify non-phrases such as "emissions vehicle." However, since they are not phrases, it is very unlikely that they appear in the document collection, so including them in a query will not change the retrieval results (in most modern, non-boolean retrieval formulas).
- In fact, the identification of non-phrases might even improve results by matching texts which contain the phrase words in high proximity, if not as a real phrase. Returning to the "emissions vehicle" example, this "phrase" will match (given the standard stemming and stopping processes common in web search) multi-word units such as "emitted from a vehicle" or "emissions of vehicles".
- Finally, this light-weight approach is extremely fast and robust, relying on no external algorithms such as linguistic analyses and statistics.

In addition to the expansion of queries with phrases, we take a similar approach to expand queries with *proximity terms*. These are also multi-word terms; the difference between them and phrase terms is that they are permitted to match not only documents which contain the phrase "as-is", but any document which contains all the words in the term inside a window of K words (K is a parameter, values are 5-15).

Experiments

To test our hypothesis, we used 125 informational queries from the web retrieval evaluation at TREC 2003 and 2004; the average length of the queries was 2.4 words (which is the same length as reported by major commercial search engines). We chose informational queries rather than navigational queries both because the performance on the latter is very high already, and does not require additional substantial work, and because phrases that appear in navigational queries tend to be proper names; as such, they are unlikely to appear in any form except as phrases, and converting them to phrases will not change the retrieval performance. The document collection we used was the TREC .GOV collection — a crawl of the .gov domain consisting of 18GB of text in 1.25 million HTML documents.

We tested our phrase-expansion method on two versions of the collection: in the first version, documents were treated as plain-text documents, meaning that all the fields (title, anchor text, etc,) were combined into a single representation of the document. In the second version, we separated the content in the documents according to the following fields: title, URL, body, and anchor text. On the first version of the collection we observed, like many others before us, no apparent gain from using phrases: sometimes a small increase in performance, sometimes a small decrease. However, on the second version we observed substantial improvements of up to 23% — depending on various parameters and the metric measured. The expansion was especially helpful for queries of length 2–3, which constitute the vast majority of all informational queries.

Conclusions

We set out to examine the usage of phrases in the domain of web retrieval. Although past research on retrieving phrases has generally not shown substantial improvements, we hypothesized that for HTML documents this may be different. The existence of short, phrase-rich, highly descriptive fields (such as title and anchor text) in these documents suggested that using phrases for their retrieval can provide a better outcome than using phrases for plain-text retrieval. Our experiments support this hypothesis, showing that the usage of phrase expansion — even with a very simple phrase detection method — substantially improves retrieval effectiveness for web documents.

Gilad Mishne received his undergraduate degree from the Technion (Israel Institute of Technology), and after a number of years in industry returned to the academic world to get an M.Sc. from the University of Amsterdam, where he is currently pursuing his Ph.D. Gilad's research is focused on Information Retrieval in the emerging domain of blogs. He can be contacted at gilad@science.uva.nl.

Acknowledgements: This paper is a summarized version of "Boosting Web Retrieval with Query Operators", appearing in: D.E. Losada and J.M. Fernández-Luna, editors, Advances in Information Retrieval: Proceedings 27th European Conference on IR Research (ECIR 2005), LNCS 3408, Springer, pages 502-516, 2005.

Both authors were supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001. Maarten de Rijke was also supported by NWO under project numbers 365-20-005, 612.069.006, 612.000.106, 612.000.207, 612.066.302, 264-70-050, and 017.001.190.