Introduction to Search Results Clustering[1]

Dawid Weiss[2]

Abstract

This paper is an introduction to the problem of search results clustering. SRC can be considered part of Web Mining, a dynamically growing branch of Information Retrieval. In this paper we give a definition of the problem, and compare it to classical document clustering. We summarize the existing body of knowledge in the field and discuss the problems with existing algorithms when applied to the Polish language. Results from an experimental clustering system are presented and problems are highlighted. Finally we propose some future research directions, listing open issues with existing algorithms.

1. Facing the Internet Search problem

A scientist faced with a new problem should do one thing before doing anything else: fully understand it. Not the solution, not even the path perhaps leading to the solution, but the problem itself. Looking at the past, we might observe that good problem understanding is the key to its proper definition and halfway to finding the correct answer. Let us introduce the problem then.

Search results clustering is a relatively new issue compared to other fields of information retrieval. Although derived from full-text classification/ grouping methods, it has several key differences, which make its definition more difficult and promote it to being a separate scientific problem worth further investigation on its own rights.

The source of SRC lies in the uncontrolled mass of information stored in the global network, the Internet. Data in the Internet is never in a consistent state – new resources are added, stale information is removed or updated. More and more powerful search engines attempt to overcome this process, constantly indexing new pages, caching the removed sites, allowing users to *search* for specific phrases or keywords and putting some structure to this big uncontrolled mess called the Web. Recent analyses show that search engines actually lose the race of keeping the information they store up to date with the actual content of the Internet [1]. Yet, nobody complains because having a broken tool is still better than having no tool at all.

Currently available search engines suffer from many things; lack of natural language-expressed queries, lack of searching focused on the meaning of terms as opposed to strict keyword matching, most engines do not even care about inflection rules of a particular language, making it very hard for users to clearly express their needs. But above all, search engines suffer from *the ranked list presentation* paradigm. To many people, it is now unthinkable to have the results displayed in any other form but a ranked list, where documents matching "best" are on top, and those less relevant are somewhere down the list. Ranking algorithms for arranging search results linearly are very often

---

[1] This is an extended abstract of a paper submitted to the 6th International Conference on Soft Computing and Distributed Processing, Rzeszów, Poland. A number of topics have been merely outlined – these were broadened during the presentation at the conference.

[2] Affiliation: Institute of Computing Science, Poznań University of Technology, Poznań, Poland, dawid.weiss@cs.put.poznan.pl

brilliant (like Google's PageRank [2]), but they preserve solely the relevance of a document to the query, losing all other profitable aspects of the results the user could utilize.

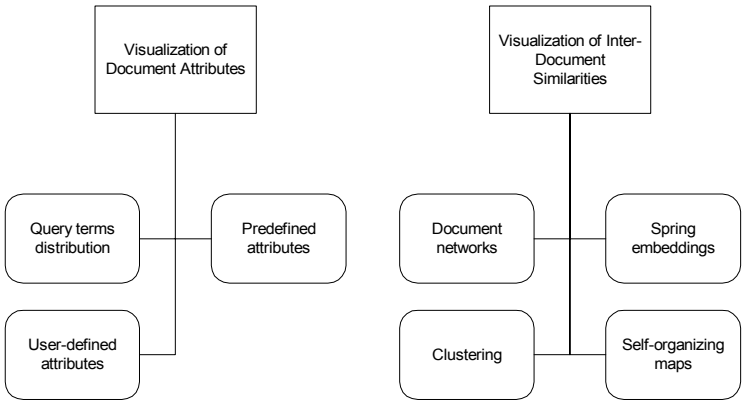

Figure 1. Other than query relevance aspects of search results visualization.

Search Results Clustering attempts to extract and display inter-document similarities within the results of a particular query, thus avoiding the danger of comparing documents concerning different subjects (such documents are most likely not comparable at all). For instance, a query "data about mining" would bring up the documents concerning *data mining*, but should also return documents about mining in general, perhaps annual summaries of the mining industry. How can documents from these two domains be put into the same ranking list? They cannot. SRC would attempt to identify the topics present in the results and compare (sort) only documents from within the same *cluster*. This is extremely useful if there is one aspect of the query (such as *data mining* in the example), which greatly outnumbers and dominates the other ones. SRC has the ability to show documents from an intersection of all the different result types, while a ranked list would display only results from the dominant cluster.

Having understood the problem, we may define it more formally:

- Let there be a certain number *N* of *search results* as returned by a traditional search engine in response to some *query*.
- Each search result represents a document in the Internet and is composed of the document's URL, an optional title and short text relevant to the document's contents, called a *snippet*. We may assume that the search engine's algorithms work in such way, that snippets are descriptive about the topic of the documents they represent, even if they are not part of those documents' body. For some documents snippets may not exist at all, or be short and form no valid language constructs (sentences).
- Let there be *P* possibly overlapping *topics* in the set of documents in the result. One document may belong to more than one topic. Topics are *semantical groups*, i.e. they represent the real *meaning* of the documents, not necessarily statistical distribution of terms (although this may be true). The set of topics may contain relations - in particular, hierarchical dependencies may exist.

The problem of Search Results Clustering may be defined by:

- Identifying the structure of *topics* and creating *clusters* representing one or more topics, if their meaning can be considered close enough. If a hierarchical structure of *topics* can be identified, it should be mapped to the arrangement of clusters.

- The algorithm must work in a reasonably short time, so that the process of clustering is transparent to the user. Incremental and linear approaches are preferred over non-polynomial ones.
- *P* should be lower than *N*, because the objective of the algorithm is to identify documents sharing common topics, thus reducing the size of the results, not producing an alternative view over them.
- The algorithm must perform well with limited size of input data (or short snippets), or indicate that it is not possible to determine any logical structure of topics in the result set.
- The algorithm should be able to *describe* the topics forming a *cluster* in a manner intuitive for the user.

1.2. Why is SRC different to document clustering?

A question arises, what is different about SRC if compared to classical document clustering or classification? The main distinction, which must be stressed here, is in assessment of algorithm's quality. Classical methods are usually evaluated based on mathematical proofs of correctness plus recall and coverage on test data. SRC is a fully user-oriented field and is evaluated by *subjective* assessment of the quality of produced clusters. More formal methods should be elaborated with time, but users' judgment about usefulness of produced clusters should always be the most important voice. In other words, any algorithm (exact or heuristic) leading to good results is considered good, no matter if it can be mathematically justified.

2. SRC body of knowledge

Search results clustering is a relatively fresh and narrow field of science. Therefore not many papers were devoted specifically to this subject. A number of applications of classical clustering methods were proposed to deal with the problem – K-means and various AHC methods just to name a few, yet they turned out not very useful in this context [3].

The first break-through algorithm in SRC was presented in *Grouper* system [3, 4]. *Suffix Tree Clustering* (STC) was an online, incremental method of finding clusters based on recurring phrases in snippets. The algorithm has several drawbacks, however, especially when it is applied to languages with a less strict sentence word order. A new approach was introduced recently – SHOC (paper to appear), supposedly overcomes STC's limitations, yet its authors give very limited number of examples and evaluation results, so it is difficult to fully assess the value of this algorithm.

Interestingly, the most advanced and the best method is currently employed in a commercial clustering search engine, Vivisimo (*www.vivisimo.com*). Not much is known about the algorithm, but the results of this service are so good, that it can be considered a benchmark and state-of-the-art in current research.

2.1 Language-aware clustering

SRC algorithms attempt to address the issue of finding *meaning* and *order* in of a set of search result, without employing any advanced language processing methods. The tradeoffs of such approach require the algorithm to work on some statistical information about frequency of words and/ or phrases. Unfortunately, human languages have a great number of constructs which distort and sometimes completely ruin the statistical distribution of frequency of terms/ phrases: synonyms, intentional altering of word order, using pronouns to refer to words once used in the text etc. These factors greatly affect the quality of clusters. For this reason, the need for language-aware algorithms evolved, where algorithms, while still not employing complex NLP methods, would utilize basic knowledge about grammar and constructs of a particular language.

3. Application of STC to Polish

At Poznań University of Technology, we developed an implementation of STC oriented toward clustering search results in Polish to empirically determine how it behaves when applied to language of more complex than English syntax. As expected, the results were of average quality and usefulness. Carrot system proved the following:

- Stop words *must* be accurately identified, otherwise they appear as junk clusters in the result,
- The standard word-weighting formula *tfidf* requires a good stemming algorithm to be used, otherwise STC applied to languages with complex flection suffers because it cannot properly determine keywords of a document.
- Order of words in phrases may not play the key role in finding base clusters. Polish is not as sensitive to the order of words as English, it is even discouraged to use repetitions of the same construct because of stylistic reasons. Because phrases are the core of STC, poor results can be expected.

There are still numerous problems and issues with application of clustering algorithms to Polish. SHOC claims to be a universal approach, applicable to any language, but it has not been tested by us as of the time of writing this paper.

4. Future research directions

Because so little has been done in SRC, there are many research directions to follow and many bold ideas waiting to be implemented. The obvious goal should be to improve the quality of clusters, their *informative* value to the user of a search engine. A very important issue here is to find some objective measure of user satisfaction, otherwise it will be hard to asses the field's progress.

Even if new algorithms are hard to come up with, existing methods have a number of drawbacks one can work on. For example, creating a better cluster merging method in STC is necessary, also further experiments on using bags of words instead of ordered phrases seem very promising.

In the light of search engines' popularity regardless of their mentioned drawbacks, SRC has a key role to play: bringing new quality tools and facilitating searching for information in the Internet. This mission can be accomplished only with proper understanding of the problem and new, specialized algorithms.

References

[1] Lawrence S., Giles L.: *Accessibility and Distribution of Information on the Web*, Nature, vol. 400, http://wwwmetrics.com

[2] Page L. et al.: The PageRank Citation Ranking: Bringing order to the Web. [@:] http://www-db.stanford.edu/~backrub/pageranksub.ps

[3] Zamir O.: Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results, Doctoral dissertation, University of Washington, 1999.

[4] Zamir O., Etzioni O., Madani O., Karp R. M.: Fast and Intuitive Clustering of Web Documents, 1997, [@:] http://citeseer.nj.nec.com/article/zamir97fast.html.