Web search results clustering in Polish: experimental evaluation of Carrot

Dawid Weiss and Jerzy Stefanowski

Institute of Computing Science, Poznań University of Technology, Poznań, Poland

Abstract. In this paper we consider the problem of web search results clustering in the Polish language, supporting our analysis with results acquired from an experimental system named Carrot. The algorithm we put into consideration – *Suffix Tree Clustering* has been acknowledged as being very efficient when applied to English. We present conclusions from its experimental application to Polish, indicating fragile areas, where the algorithm seem to fail due to specific properties of the input data. We indicate that the characteristics of produced clusters (number, value), unlike in English, strongly depend on pre-processing phase. We also attempt to investigate the influence of two primary STC parameters: *merge threshold* and *minimum base cluster score* on the number and quality of results. Finally, we introduce two approaches to efficient, approximate stemming of Polish words: *quasi-stemmer* and an automaton-based method.

1 Search results clustering overview

Together with an exponential increase of the number of documents available in the internet, comes the requirement for faster and more reliable tools for locating relevant information - the role currently played by internet search engines. While algorithms for indexing and querying large volumes of data have been substantially improved, the paradigm of searching for information based on providing query terms and retrieving a list of matching documents, remained almost the same since the very beginning. Considering solely the number of matching results for even very narrow topics, browsing such set is not feasible anymore. Besides, query terms are often ambiguous and spanning over multiple subjects, making it impossible to present the results as a linear ordering of relevance.

Searching and browsing are in very strong relationship to each other; it is natural that improving browsing techniques will also improve the overall performance of seeking for information. Search engines return a list of references to matching documents, each one usually comprising a title, URL and a short fragment of the source document, called a *snippet*. A snippet should contain enough information about the document it describes, to give the user a clue what the entire document is about. This is the starting point for search results clustering, which attempts to form *meaningful, thematic groups* of snippets. These groups are then presented instead, or in addition to the original document references. The user gets a much deeper insight into

the subjects the query covered, simply by looking at the set of discovered groups.

2 STC algorithm and related work

The problem of clustering in general is very well explored, thanks to the heritage drawn from statistics, economy and even fields of computer science such as data mining and information retrieval. In spite of this, several properties such as demand for linear complexity, incomplete input data (a snippet instead of full body of a document) and processing of text features render search result clustering a research field worth investigating on its own.

Arguably the first query result visualization algorithm based on the paradigm of clustering was presented in Scatter-Gather system [1], but it was not until *Suffix Tree Clustering* (STC) technique appeared [7], that the field of search results clustering, also called *ephemeral clustering* [4] had been given a substantial momentum.

STC is an algorithm with at least two distinguishing features: its time complexity is linear with respect to the number of clustered snippets, and it operates on phrases present in the text, in contrast to most previous efforts, built on top of standard IR measures of terms frequency distribution. STC attempts to cluster documents or search results according to shared *phrases* they contain, thus employing information about the proximity of terms, in addition to their frequencies.

STC is organized into two phases: discovering *base clusters* and combining (merging) them to form the final set of groups. In the first phase, all sentences building the search result (snippets and document titles) are inserted into a *generalized suffix tree*, where each symbol (node) in the tree represents a single term. Every node holds references to the sentences (and documents) it occurred in. Thus, each path from some node to the root of the suffix tree denotes a phrase shared by all the documents the node holds references to. For each phrase shared by at least two documents, we define a *base cluster score*: $s(m) = |m| \times f(|m_p|) \times \sum(tfidf(w_i))$, where $|m|$ is the number of non-stopped terms in phrase m, $f(|m_p|)$ is a function penalizing short-phrases, $tfidf(w_i)$ is a standard Salton's *term frequency-inverse document frequency* term ranking measure. A set of base clusters is identified by selecting phrases with base cluster score higher than an arbitrarily chosen *minimal base cluster score threshold*. In the second stage of the algorithm N top-ranking base clusters are merged using a version of AHC algorithm, with binary single-link merge criterion between base clusters a and b defined by the formula: $similarity(a,b) = 1 \Leftrightarrow \left(\frac{|a \cap b|}{|a|} > \alpha\right) \wedge \left(\frac{|a \cap b|}{|b|} > \alpha\right)$, 0 otherwise, where α in the above formula denotes an arbitrarily chosen *merge threshold*.

Unfortunately due to space constraints, we are not able to give a broader insight into algorithms that followed STC (refer to [4] for a review of existing

methods), but they all had one common drawback: were designed, implemented and evaluated for English only. This puts in question their applicability to other languages, because, as we are about to show, the properties of a language severely affect algorithm's performance. To our best knowledge only Semantic Hierarchical Online Clustering (SHOC) algorithm [8] attempts to overcome the language issues. SHOC was designed to work in Chinese and its authors claim it works very well, however it is not available for comparisons.

Stunningly, with abundance of new algorithms, there exists a significant shortage in evaluation techniques and measures of ephemeral clustering quality. Most algorithms are judged based on explicit user surveys or analysis of server logs, comparing raw search interface to the clustered one [7], in spite of the fact that [5] clearly states such comparison can be misleading. A more objective, information entropy – derived measure is used for evaluation in [4]. We chose to employ this one as well, even though it has certain shortcomings discussed in section 4.

3 Motivation

It seems that the issue of language-aware search results clustering is to some point neglected. We decided to investigate the behavior of STC, one of the most widely known algorithms, specific to the problem, and decided to investigate various aspects of its performance when applied to Polish. In particular, we wanted to examine the impact, which inflection, meaningless terms (stop-words), word order, etc. have on the quality of results returned by STC.

This was one of the main reasons of creating Carrot system. Carrot is an implementation of STC, with extensions such as stemming and stop-words, specifically designed for use with English and Polish queries. Carrot was available online between Spring and Summer of 2001, before legal issues concerning automated querying of the background major search company forced us to limit its public availability[1].

Another reason driving this experiment was to find out how the choice of STC's thresholds affect the quality and stability of results. We also wanted to employ a mathematical measure of quality as opposed to user surveys used in [7].

4 The experiment

In order to fulfill the goals given above, we prepared a small-scale experiment comparing clusters acquired from Carrot to a predefined, manually produced grouping. In order to keep the experiment realistic, we decided to utilize two sets of real search results downloaded from Vivisimo search engine in response

[1] More information can be found at *http://www.cs.put.poznan.pl/dweiss/carrot*

to queries: *inteligencja* (intelligence) and *odkrywanie wiedzy* (knowledge discovery). It should be stressed that only these two test queries were used for the experiment, but the longer-term experiences with Carrot seem to support the results we obtained. We realize the fact that our experimental data should be larger – it is a matter of further research, yet we decided to publish the results of this preliminary experiment, because we think it clearly illustrates the deficiencies STC seems to have.

The manual clustering of the two test queries had been performed by five experts in the field, independently. Original documents were not retrieved from the Web, so that humans had as much information about the clustered set, as the algorithm – a set of about 70 snippets per query. This again proved that clustering, even when done by hand, is unambiguous and problematic (ref. to [5]). Out of 80 snippet-to-cluster assignments, only 50% were fully consistent among all individuals. There were about 14 manually created groups for each query.

Having a manual benchmark to compare against, we faced the problem of choosing possibly objective comparison function. The problems with empirical evaluation of ephemeral clustering systems have already been mentioned in section 2. We decided to utilize Byron Dom's method [2], briefly presented in definition 1. The manual clustering we produced was used as the *ground truth* set required by the measure.

The experiment consisted of comparing results produced by STC to the ground truth set using the measure given in definition 1. Full range of values for the key algorithm parameters – merge threshold and minimal base cluster score (see section 2 for definition) – were taken into account. The experiment was repeated for different configurations of data pre-processing – with and without stemming and stop words filtering, which will be explained further.

Definition 1. Let X be a set of *feature vectors*, representing objects to be clustered, C be a set of *class labels*, representing the desired, optimal classification (also called a *ground truth set*), and K be a set of cluster labels assigned to elements of X as a result of an algorithm.

Knowing X and K one can calculate a two-dimensional contingency matrix $H \equiv \{h(c,k)\}$, where $h(c,k)$ is the number of objects labeled class c that are assigned to cluster k. *Information-theoretic external cluster-validity measure* is defined by:

$$Q_0 = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{h(c,k)}{n}\log\frac{h(c,k)}{h(k)} + \frac{1}{n}\sum_{k=1}^{|K|}\log\binom{h(k)+|C|-1}{|C|-1}. \quad (1)$$

where $n = |X|$, $h(c) = \sum_k h(c,k)$, $h(k) = \sum_c h(c,k)$.

For the experiment, we used a normalized version of the formula presented in definition 1, where zero denotes no correspondence between ground truth set of clusters and the compared set, while one means the two are identical.

Byron Dom's measure is defined for flat partitioning of an input set of objects, while STC produces flat, but overlapping clusters. In order to be able to apply the measure, we decided that snippets assigned more than one cluster, would belong only to the one having the highest score assigned by STC. Obviously, this approach favors large, strong clusters, but we found no other, objective pattern of making STC's result conform to the assumptions of the measure.

4.1 Stemming algorithms used for experiments

According to our knowledge there are no publicly available Polish stemming algorithms. As part of our work on Carrot, we created two efficient, approximate stemming methods for the Polish language.

Quasi-stemmer is based on the use of over eight hundred thousand unique terms corpora of Polish texts, obtained by permission from Rzeczpospolita newspaper archives. About five hundred most common suffixes from this corpora have been extracted and put in a lookup-table. The condition for determining whether two words α and β can be considered inflected forms of a common stem is given in definition 2.

Definition 2. Two words α and β may be considered to originate from a common stem, if they contain a common prefix, which is at least n characters long, and the suffixes created by stripping that prefix exist in the predefined lookup table.

It is obvious that such simple technique has limited accuracy, hence the prefix *quasi-* in the name. Also, the equivalence relation implied by quasi-stemmer is not transitive. We avoided this problem by comparing not terms themselves, but building equivalence classes. Once a term has been assigned to a class, it cannot be reassigned to another.

Quasi-stemmer brings substantial improvement in both cognitive and experimental results (refer to section 5.1).

Driven by desire for even higher quality stemming, we created another technique, this time based on an open source dictionary of Polish included with `ispell`[2]. All the possible inflected forms of each word were generated first, utilizing ispell's information about flection rules. Thus, for each term we had a mapping to its stem. We used finite state automatons to compress this information into a compact data structure used directly for lookups. This approach proved to be superior to quasi-stemmer both in speed and quality. For any term, the complexity of its stem-lookup operation is at most linear with the length of the term (this only, if the lookup has been successful).

[2] *http://ispell-pl.sourceforge.net*

5 Results

5.1 Inflection and stop words

Polish inflection rules are much more complex than English. Words have different suffixes depending on the case, gender, number, person, degree, mode or tense. Suffixes are mostly regular, but many exceptions exist and construction of a straightforward rule-based stemmer, like Porter stemmer for English, is not an easy task.

Oren Zamir in [7] claims that the importance of preprocessing phase of STC, involving stemming and removal of stop words, is not a crucial factor of algorithm's performance. Our experiments and papers, such as [3], clearly indicate it is not the case. We discovered that both cognitive and measured results improve substantially with the use of input data preprocessing. This involves both discarding stop words and stemming. In figure 1 one can clearly see that for both analyzed queries, the highest scoring configuration was with the use of stop words and the dictionary-based stemmer.

While the improvement brought by quality stemming can be easily explained (refer to [3]), it is quite unexpected that stop words have such high influence on results (even though preliminary step of STC discards the most frequent terms occurring in the input). This suggests that a more advanced reduction of meaningless terms could drastically improve cluster quality. Maybe this reduction could even go as far as discarding all non-nouns, or non-verbs.

Inflected words also have a negative influence on the factor used for calculating cluster score. The TFIDF (*term frequency inverse document frequency*) model used for scoring terms in phrases does not yield the actual importance of a given word, because the score is distributed over all of its inflected forms.

5.2 Phrases versus words

STC produces clusters based on common phrases occurring in snippets. The Polish language has a feature of not being as order-dependant as English. The meaning of a sentence can be as well carried solely by suffixes of the terms used. Compare this classic example from [6]: *John hit Paul* $\neq$ *Paul hit John*, while in Polish: *Jan uderzył Pawła* = *Pawła uderzył Jan.* Consequently, STC's primary assumption that common phrases form good clusters is very problematic. Unfortunately, even stemming can cause troubles in this example, because if words are brought back to their stems, we are no longer able to tell who hit whom: *Jan uderzył Pawła* $\neq$ *Jana uderzył Paweł*, (both of which stem to) *Jan uderzać Paweł*. The only solution to this problem would be processing of semantics. While this is rather a long-term perspective (because of speed), we propose using n-grams instead of phrases as a research direction. This has been investigated for English by [7], yielding no improvement over phrases, but in Polish this may no longer be the case.

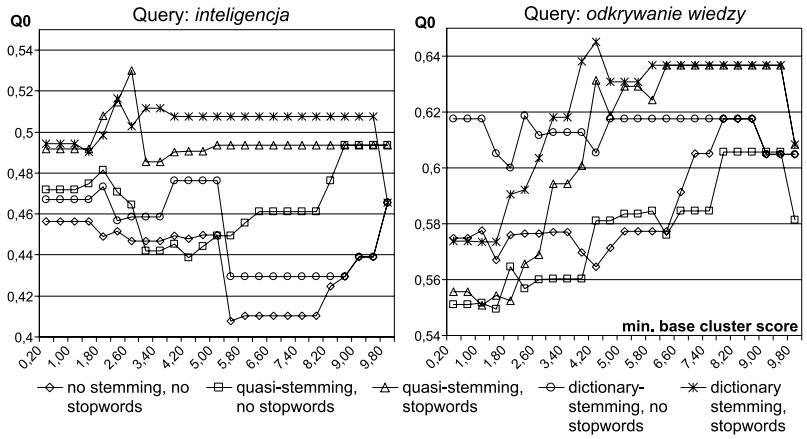

Fig. 1. Value of quality measure Q_0 in relation to changing base cluster size. Five different configurations of input preprocessing are shown. STC merge threshold is constant (0.6)

5.3 Sensitiveness of STC control parameters

Oren Zamir in [7] claims that the influence of merge threshold over the results was unnoticeable and chooses an arbitrary value for his STC evaluation. We analyzed the results acquired from all experimental configurations, trying to find arguments either to support or to refute this claim.

We have found that the results are not very sensitive to the setting of merge threshold, only when the input data has been pre-processed. Figure 2 illustrates that for raw input even a small change of the merge threshold results in completely different value of quality measure.

Minimum base cluster score threshold expresses much stronger influence, especially on the number of discovered clusters. While this is obvious (because this parameter describes a cut-off threshold for phrases considered at later stages of STC), it is not at all clear what exact value this parameter should be set to. It seems that higher values stabilize the properties of merge threshold, also improving numerical measure of quality, but it also means that low-scoring clusters, perhaps interesting, are discarded in favor of large, strong ones (which are potentially obvious to the user). Also, as seen in figure 2, when pre-processing has been applied, there is a quick saturation of quality and a sudden drop in the number of clusters produced. The results are therefore very sensitive to the exact setting of base cluster cut-off threshold. The relation between the number of clusters, their quality and the minimum base cluster score should be a matter of further research, because it severely affects the results of STC.

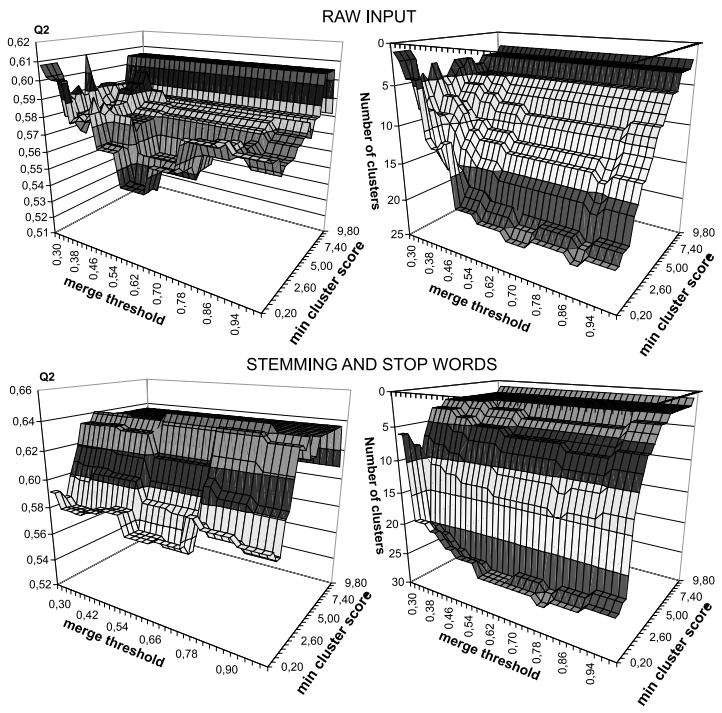

Fig. 2. A three-dimensional view of the relation between merge threshold and minimum base cluster score to the value of quality measure Q_2 (charts on the left side) and number of produced clusters (charts on the right side). Please note the axis of number of clusters is reversed for clarity

6 Conclusions and future research directions

In this paper we have presented conclusions drawn from an experimental application of STC to documents in Polish. The following fragile areas of STC algorithm have been observed: STC seems to be very sensitive to complex inflection forms in the input. This affects both phrase construction and scoring of terms. High quality stemming is therefore a necessity, not an option.

The use of phrases for finding base clusters may not necessarily be the best choice, as argued in [7]. Word order in Polish may be tricky, and the same meaning may be expressed without using exactly the same sequence of terms. The most promising research direction in this area includes verifying whether the use of n-grams as opposed to phrases would improve quality of STC.

Filtering of meaningless terms in pre-processing phrase has a significant impact on the stability and quality of clusters. An assessment of the usefulness of some advanced technique for determining these terms would be valuable.

This paper also introduces two efficient methods for stemming Polish texts, and shows how they improve the quality of results when applied to search results clustering. Potential research directions in this area should be directed at developing stemming methods capable of correctly processing proper names. Quasi-stemmer presented by us in this paper has this property to some extent, but again, a comparison of an advanced method influence over results would be very interesting.

Acknowledgements

The authors would like to thank Mr. Stanisław Osiński, Mr. Paweł Kowalik and Mr. Michał Wróblewski for their help in manual clustering.

References

1. Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. (1992) Scatter/Gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval
2. Dom E. B. (2001) An information-theoretic external cluster validity measure. IBM research report RJ 10219
3. Kantrowitz M., Behrang M., Mittal V. (2000) Stemming and its effects on TFIDF Ranking. Proceedings of the 23rd annual international ACM SIGIR, Athens, Greece
4. Maarek Y., Fagin R., Ben-Shaul I., et al. (2000) Ephemeral Document Clustering for Web Applications. IBM research report RJ 10186
5. Macskassy S. A., Banerjee A., Davison B. D., Hirsh H. (1998) Human Performance on Clustering Web Pages: A Preliminary Study. The Fourth International Conference on Knowledge Discovery and Data Mining, New York
6. Szpakowicz S. (1978) Automatyczna analiza składniowa polskich zdań pisanych. Doctoral dissertation, Warsaw University
7. Zamir O. (1999) Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results, Doctoral dissertation, University of Washington
8. Zhang D., Dong Y. (2001) Hierarchical, Online Clustering of Web Search Results, accepted by 3rd International Workshop on Web information and data management, Atlanta, Georgia