Easy Web Search Results Clustering: When Baselines Can Reach State-of-the-Art Algorithms

Jose G Moreno, Gaël Dias

To cite this version:

Jose G Moreno, Gaël Dias. Easy Web Search Results Clustering: When Baselines Can Reach State-of-the-Art Algorithms. 14th Conference of the European Chapter of the Association for Computational Linguistics, Apr 2014, Gotenburg, Sweden. 2014. <hal-01076535>

HAL Id: hal-01076535

https://hal.archives-ouvertes.fr/hal-01076535

Submitted on 22 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Easy Web Search Results Clustering: When Baselines Can Reach State-of-the-Art Algorithms

Jose G. Moreno
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
jose.moreno@unicaen.fr

Gaël Dias
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
gael.dias@unicaen.fr

Abstract

This work discusses the evaluation of baseline algorithms for Web search results clustering. An analysis is performed over frequently used baseline algorithms and standard datasets. Our work shows that competitive results can be obtained by either fine tuning or performing cascade clustering over well-known algorithms. In particular, the latter strategy can lead to a scalable and real-world solution, which evidences comparative results to recent text-based state-of-the-art algorithms.

1 Introduction

Visualizing Web search results remains an open problem in Information Retrieval (IR). For example, in order to deal with ambiguous or multifaceted queries, many works present Web page results using groups of correlated contents instead of long flat lists of relevant documents. Among existing techniques, Web Search Results Clustering (SRC) is a commonly studied area, which consists in clustering “on-the-fly” Web page results based on their Web snippets. Therefore, many works have been recently presented including task adapted clustering (Moreno et al., 2013), meta clustering (Carpineto and Romano, 2010) and knowledge-based clustering (Scaiella et al., 2012).

Evaluation is also a hot topic both in Natural Language Processing (NLP) and IR. Within the specific case of SRC, different metrics have been used such as $F_1$-measure ($F_1$), $kSSL$\footnote{This metric is based on subjective label evaluation and as such is out of the scope of this paper.} and $F_{\beta_3}$-measure ($F_{\beta_3}$) over different standard datasets: ODP-239 (Carpineto and Romano, 2010) and Moresque (Navigli and Crisafulli, 2010). Unfortunately, comparative results are usually biased as baseline algorithms are run with default parameters whereas proposed methodologies are usually tuned to increase performance over the studied datasets. Moreover, evaluation metrics tend to correlate with the number of produced clusters.

In this paper, we focus on deep understanding of the evaluation task within the context of SRC. First, we provide the results of baseline algorithms with their best parameter settings. Second, we show that a simple cascade strategy of baseline algorithms can lead to a scalable and real-world solution, which evidences comparative results to recent text-based algorithms. Finally, we draw some conclusions about evaluation metrics and their bias to the number of output clusters.

2 Related Work

Search results clustering is an active research area. Two main streams have been proposed so far: text-based strategies such as (Hearst and Pedersen, 1996; Zamir and Etzioni, 1998; Zeng et al., 2004; Osinski et al., 2004; Carpineto and Romano, 2010; Carpineto et al., 2011; Moreno et al., 2013) and knowledge-based ones (Ferragina and Gulli, 2008; Scaiella et al., 2012; Di Marco andNavigli, 2013). Successful results have been obtained by recent works compared to STC (Zamir and Etzioni, 1998) and LINGO (Osinski et al., 2004) which provide publicly available implementations, and as a consequence, are often used as state-of-the-art baselines. On the one hand, STC proposes a monothetic methodology which merges base clusters with high string overlap relying on suffix trees. On the other hand, LINGO is a polythetic solution which reduces a term-document matrix using single value decomposition and assigns documents to each discovered latent topic.

All solutions have been evaluated on different datasets and evaluation measures. The well-known $F_1$ has been used as the standard evaluation metric. More recently, (Carpineto and Romano,
Different standard datasets have been built such as AMBIENT$^2$ (Carpineto and Romano, 2009), ODP-239$^3$ (Carpineto and Romano, 2010) and Moresque$^4$ (Navigli and Crisafulli, 2010). ODP-239, an improved version of AMBIENT, is based on DMOZ$^5$ where each query, over 239 ones, is a selected category in DMOZ and its associated sub-categories are considered as the respective cluster results. The small text description included in DMOZ is considered as a Web snippet. Moresque is composed by 114 queries selected from a list of ambiguous Wikipedia entries. For each query, a set of Web results have been collected from a commercial search engine and manually classified into the disambiguation Wikipedia pages which form the reference clusters.

In Table 2, we report the results obtained so far in the literature by text-based and knowledge-based strategies for the standard $F_1$ over ODP-239 and Moresque datasets.

Table 2: State-of-the-art Results for SRC. (*) The result of (Di Marco and Navigli, 2013) is based on a reduced version of AMBIENT + Moresque.

<table>
<thead>
<tr>
<th>Text</th>
<th>$F_1$ ODP-239</th>
<th>$F_1$ Moresque</th>
</tr>
</thead>
<tbody>
<tr>
<td>STC</td>
<td>0.324</td>
<td>0.455</td>
</tr>
<tr>
<td>LINGO</td>
<td>0.272</td>
<td>0.266</td>
</tr>
<tr>
<td>Moresque (Carpineto and Romano, 2010)</td>
<td>0.315</td>
<td>0.362</td>
</tr>
<tr>
<td>Moresque (Moreno et al., 2013)</td>
<td>0.390</td>
<td>0.665</td>
</tr>
<tr>
<td>(Scarcelli et al., 2012)</td>
<td>0.413</td>
<td>-0.7204*</td>
</tr>
<tr>
<td>(Di Marco and Navigli, 2013)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In the previous section, our aim was to claim that tunable versions of existing baseline algorithms might evidence improved results when faced to the ones reported in the literature. And these values should be taken as the “real” baseline results within the context of controllable environments. However, exploring all the parameter space is not an applicable solution in a real-world situation where the reference is unknown. As such, a stopping criterion must be defined to adapt to any dataset distribution. This is the particular case for the standard implementations of STC and LINGO.

**4 Cascade SRC Algorithms**

In Table 1, we report the results obtained so far in the literature by text-based and knowledge-based strategies for the standard $F_1$ over ODP-239 and Moresque datasets.

Table 1: Standard, Tuned and Random Results for Moresque and ODP-239 datasets.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$F_1$ ODP-239</th>
<th>$F_1$ Moresque</th>
</tr>
</thead>
<tbody>
<tr>
<td>STC</td>
<td>0.324</td>
<td>0.455</td>
</tr>
<tr>
<td>LINGO</td>
<td>0.272</td>
<td>0.266</td>
</tr>
<tr>
<td>Moresque</td>
<td>0.315</td>
<td>0.362</td>
</tr>
<tr>
<td>Moresque</td>
<td>0.390</td>
<td>0.665</td>
</tr>
</tbody>
</table>

default parameters based on available implementations. As such, no conclusive remarks can be drawn knowing that tuned versions might provide improved results.

In particular, available implementations$^6$ of STC, LINGO and the Bisection $K$-means ($BiKm$) include a fixed stopping criterion. However, it is well-known that tuning the number of output clusters may greatly impact the clustering performance. In order to provide fair results for baseline algorithms, we evaluated a $k$-dependent$^7$ version for all baselines. We ran all algorithms for $k = 2..20$ and chose the best result as the “optimal” performance. Table 1 sums up results for all the baselines in their different configurations and shows that tuned versions outperform standard (available) ones both for $F_1$ and $F_{b3}$ over ODP-239 and Moresque.

**3 Baseline SRC Algorithms**

Newly proposed algorithms are usually tuned towards their maximal performance. However, the results of baseline algorithms are usually run with
In a cascade strategy, we first cluster the initial set of Web page snippets with any SRC algorithm. Then, the input of the second SRC algorithm is the set of meta-documents built from the documents belonging to the same cluster. Finally, each clustered meta-document is mapped to the original documents generating the final clusters. This process can iteratively be applied, although we only consider two-level cascade strategies in this paper.

This strategy can be viewed as an easy, reproducible and parameter free baseline SRC implementation that should be compared to existing state-of-the-art algorithms. Table 3 shows the results obtained with different combinations of SRC baseline algorithms for the cascade strategy both for $F_1$ and $F_{\beta_3}$ over ODP-239 and Moresque. The “Stand.” column corresponds to the performance of the cascade strategy and $k$ to the automatically obtained number of clusters. Results show that the combination STC-STC achieves the best performance overall for the $F_1$ and STC-LINGO is the best combination for the $F_{\beta_3}$ in both datasets.

In order to provide a more complete evaluation, we included in column “Equiv.” the performance that could be obtained by the tunable version of each single baseline algorithm based on the same $k$. Interestingly, the cascade strategy outperforms the tunable version for any $k$ for $F_1$ but fails to compete (not by far) with $F_{\beta_3}$. This issue will be discussed in the next section.

5 Discussion

In Table 1, one can see that when using the tuned version and evaluating with $F_1$, the best performance for each baseline algorithm is obtained for the same number of output clusters independently of the dataset (i.e. around 3 for STC and LINGO and 2 for BiKm). As such, a fast conclusion would be that the tuned versions of STC, LINGO and BiKm are strong baselines as they show similar behaviour over datasets. Then, in a realistic situation, $k$ might be directly tuned to these values.

However, when comparing the output number of clusters based on the best $F_1$ value to the reference number of clusters, a huge difference is evidenced. Indeed, in Moresque, the ground-truth average number of clusters is 6.6 and exactly 10 in ODP-239. Interestingly, $F_{\beta_3}$ shows more accurate values for the number of output clusters for the best tuned baseline performances. In particular, the best $F_{\beta_3}$ results are obtained for LINGO with 5.8 clusters for Moresque and 8.7 clusters for ODP-239 which most approximate the ground-truths.

In order to better understand the behaviour of each evaluation metric (i.e. $F_\beta$ and $F_{\beta_3}$) over different $k$ values, we experienced a uniform random clustering over Moresque and ODP-239. In Figure 1(c), we illustrate these results. The important issue is that $F_\beta$ is more sensitive to the number of output clusters than $F_{\beta_3}$. On the one hand, all $F_\beta$ measures provide best results for $k = 2$ and a random algorithm could reach $F_1=0.5043$ for Moresque and $F_1=0.2980$ for ODP-239 (see Table 1), thus outperforming almost all standard implementations of STC, LINGO and BiKm for both datasets. On the other hand, $F_{\beta_3}$ shows that most standard baseline implementations outperform the random algorithm.

Moreover, in Figures 1(a) and 1(b), we illustrate the different behaviours between $F_1$ and $F_{\beta_3}$ for $k = 2.20$ for both standard and tuned versions of STC, LINGO and BiKm. One may clearly see that $F_{\beta_3}$ is capable to discard the algorithm (BiKm) which performs worst in the standard version while this is not the case for $F_1$. And, for LINGO, the optimal performances over Moresque and ODP-239 are near the ground-truth number of clusters while this is not the case for $F_1$ which evidences a decreasing tendency when $k$ increases.

In section 4, we showed that competitive results could be achieved with a cascade strategy based on baseline algorithms. Although results outperform standard and tunable baseline implementations for $F_1$, it is wise to use $F_{\beta_3}$ to better evaluate the SRC task, based on our previous discussion. In this case, the best values are obtained by STC-LINGO with $F_{\beta_3}=0.4980$ for Moresque and $F_{\beta_3}=0.4249$ for ODP-239, which highly approximate the values reported in (Moreno et al., 2013): $F_{\beta_3}=0.490$ (Moresque) and $F_{\beta_3}=0.452$ (ODP-239). Additionally, when STC is performed first and LINGO later the cascade algorithm scale better due to LINGO and STC scaling properties.

6 Conclusion

This work presents a discussion about the use of baseline algorithms in SRC and evaluation met
Our experiments show that $F_{b3}$ seems more adapted to evaluate SRC systems than the commonly used $F_1$ over the standard datasets available so far. New baseline values which approximate state-of-the-art algorithms in terms of clustering performance can also be obtained by an easy, reproducible and parameter free implementation (the cascade strategy) and could be considered as the “new” baseline results for future works.
References


