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 multi-faceted 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), $kSSL$[1] and F_{b^3}-measure (F_{b^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 and Navigli, 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,

[1] This metric is based on subjective label evaluation and as such is out of the scope of this paper.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–5, Gothenburg, Sweden, April 26-30 2014.

	Moresque								ODP-239							
	F_1				F_{b^3}				F_1				F_{b^3}			
Algo.	Stand.	k	Tuned	k	Stand.	k	Tuned	k	Stand.	k	Tuned	k	Stand.	k	Tuned	k
STC	0.4550	12.7	0.6000	2.9	0.4602	12.7	0.4987	2.9	0.3238	12.4	0.3350	3.0	0.4027	12.4	0.4046	14.5
LINGO	0.3258	26.7	**0.6034**	3.0	0.3989	26.7	**0.5004**	5.8	0.2029	27.7	0.3320	3.0	0.3461	27.7	**0.4459**	8.7
BiKm	0.3165	9.7	0.5891	2.1	0.3145	9.7	0.4240	2.1	0.1995	12.1	**0.3381**	2.2	0.3074	12.1	0.3751	2.2
Random	-	-	0.5043	2	-	-	0.3548	2	-	-	0.2980	2	-	-	0.3212	2

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

2010) evidenced more complete results with the general definition of the F_β-measure for $\beta = \{1, 2, 5\}$, (Navigli and Crisafulli, 2010) introduced the Rand Index metric and (Moreno et al., 2013) used F_{b^3} introduced by (Amigó et al., 2009) as a more adequate metric for clustering.

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 subcategories 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.

		F_1	
		ODP239	Moresque
Text	STC	0.324	0.455
	LINGO	0.273	0.326
	(Carpineto and Romano, 2010)	0.313	-
	(Moreno et al., 2013)	0.390	0.665
Know.	(Scaiella et al., 2012)	0.413	-
	(Di Marco and Navigli, 2013)	-	0.7204*

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.

3 Baseline SRC Algorithms

Newly proposed algorithms are usually tuned towards their maximal performance. However, the results of baseline algorithms are usually run with 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_{b^3} over ODP-239 and Moresque.

4 Cascade SRC Algorithms

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.

Previous results (Carpineto and Romano, 2010) showed that different SRC algorithms provide different results and hopefully complementary ones. For instance, STC demonstrates high recall and low precision, while LINGO inversely evidences high precision for low recall. Iteratively applying baseline SRC algorithms may thus lead to improved results by exploiting each algorithm's strengths.

[2]http://credo.fub.it/ambient/ [Last acc.: Jan., 2014]
[3]http://credo.fub.it/odp239/ [Last acc.: Jan., 2014]
[4]http://lcl.uniroma1.it/moresque/ [Last acc.: Jan., 2014]
[5]http://www.dmoz.org [Last acc.: Jan., 2014]
[6]http://carrot2.org [Last acc.: Jan., 2014]
[7]Carrot2 parameters *maxClusters*, *desiredClusterCount-Base* and *clusterCount* were used to set k value.

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[8]. 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_{b^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_{b^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_{b^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_{b^3} shows more accurate values for the number of output clusters for the best tuned baseline performances. In particular, the best F_{b^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_β and F_{b^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_β is more sensitive to the number of output clusters than F_{b^3}. On the one hand, all F_β 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_{b^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_{b^3} for $k = 2..20$ for both standard and tuned versions of STC, LINGO and BiKm. One may clearly see that F_{b^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_{b^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_{b^3}=0.4980 for Moresque and F_{b^3}=0.4249 for ODP-239, which highly approximate the values reported in (Moreno et al., 2013): F_{b^3}=0.490 (Moresque) and F_{b^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[9].

6 Conclusion

This work presents a discussion about the use of baseline algorithms in SRC and evaluation met-

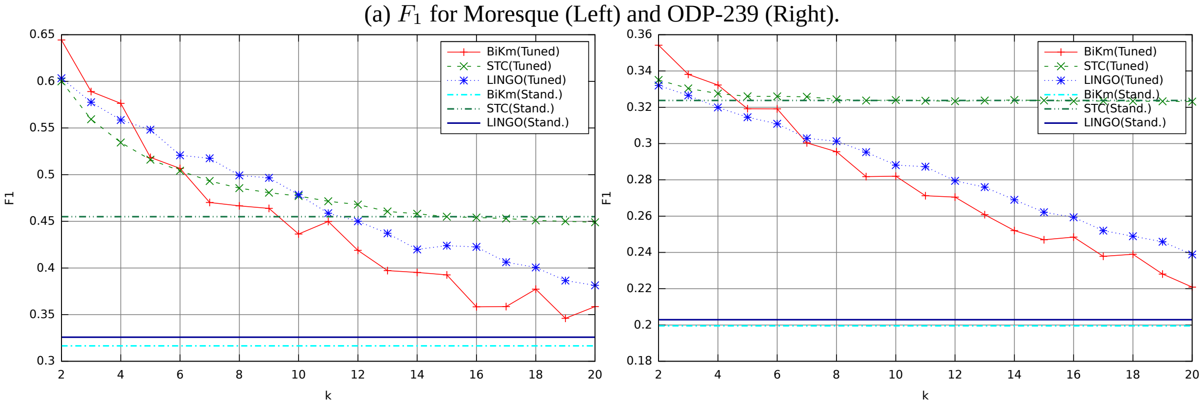
[8]Fused using concatenation of strings.

[9]http://carrotsearch.com/lingo3g-comparison [Last acc.: Jan., 2014]

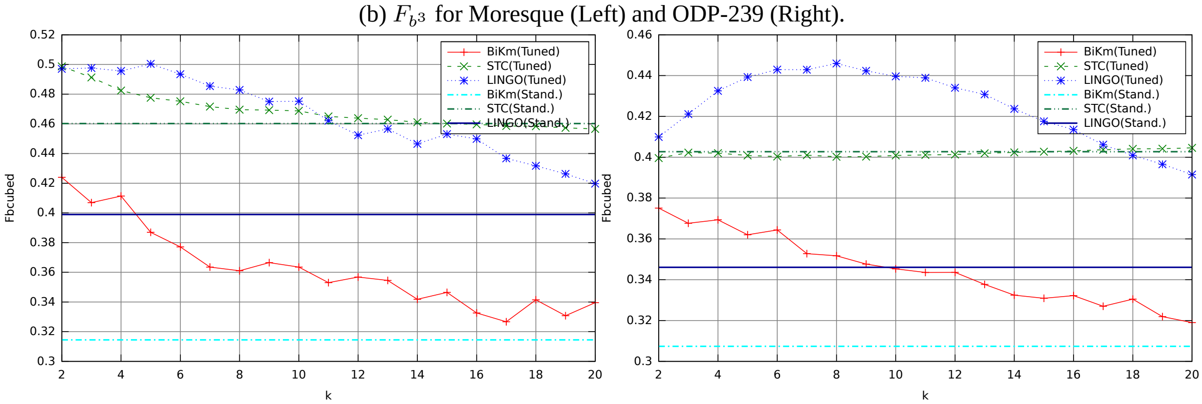
		Moresque						ODP-239					
		F_1			F_{b^3}			F_1			F_{b^3}		
Level 1	Level 2	Stand.	Equiv.	k	Stand.	Equiv.	k	Stand.	Equiv.	k	Stand.	Equiv.	k
STC	STC	**0.6145**	0.5594	3.1	0.4550	**0.4913**	3.1	**0.3629**	0.3304	3.2	0.3982	0.4023	3.2
	LINGO	0.5611	0.4932	7.3	**0.4980**	0.4716	7.3	0.3624	0.3258	6.9	**0.4249**	0.4010	6.9
	BiKm	0.5413	0.5160	4.5	0.4395	0.4776	4.5	0.3319	0.3276	4.3	0.3845	0.4020	4.3
LINGO	STC	0.5696	0.5176	6.7	0.4602	0.4854	6.7	0.3457	0.3029	7.2	0.4229	**0.4429**	7.2
	LINGO	0.4629	0.4371	13.7	0.4447	0.4566	13.7	0.2789	0.2690	13.6	0.3931	0.4237	13.6
	BiKm	0.4038	0.4966	8.6	0.3801	0.4750	8.6	0.2608	0.2953	8.5	0.3510	0.4423	8.5
BiKm	STC	0.5873	**0.5891**	2.7	0.4144	0.4069	2.7	0.3425	0.3381	2.7	0.3787	0.3677	2.7
	LINGO	0.4773	0.5186	5.4	0.3832	0.3869	5.4	0.2819	0.3191	6.3	0.3546	0.3644	6.3
	BiKm	0.4684	0.5764	3.5	0.3615	0.4114	3.5	0.2767	**0.3322**	4.3	0.3328	0.3693	4.3

Table 3: Cascade Results for Moresque and ODP-239 datasets.

(a) F_1 for Moresque (Left) and ODP-239 (Right).



(b) F_{b^3} for Moresque (Left) and ODP-239 (Right).



(c) Evaluation Metrics for Random Clustering for Moresque (Left) and ODP-239 (Right).

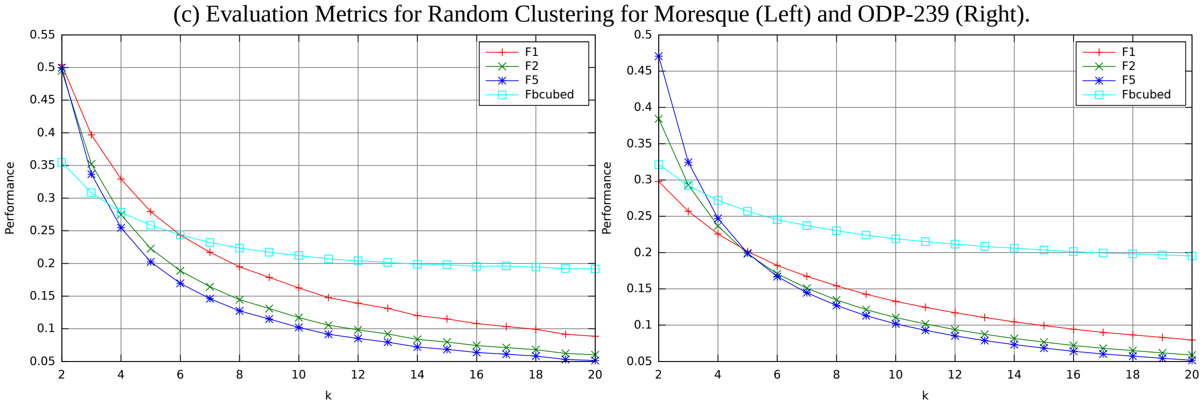


Figure 1: F_1 and F_{b^3} for Moresque and ODP-239 for Standard, Tuned and Random Clustering.

rics. Our experiments show that F_{b^3} 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

E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

C. Carpineto and G. Romano. 2009. Mobile information retrieval with search results clustering : Prototypes and evaluations. *Journal of the American Society for Information Science*, 60:877–895.

C. Carpineto and G. Romano. 2010. Optimal meta search results clustering. In *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 170–177.

C. Carpineto, M. D'Amico, and A. Bernardini. 2011. Full discrimination of subtopics in search results with keyphrase-based clustering. *Web Intelligence and Agent Systems*, 9(4):337–349.

A. Di Marco and R. Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

P. Ferragina and A. Gulli. 2008. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225.

M.A. Hearst and J.O. Pedersen. 1996. Re-examining the cluster hypothesis: Scatter/gather on retrieval results. In *19th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 76–84.

J.G. Moreno, G. Dias, and G. Cleuziou. 2013. Post-retrieval clustering using third-order similarity measures. In *51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 153–158.

R. Navigli and G. Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 116–126.

S. Osinski, J. Stefanowski, and D. Weiss. 2004. Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Systems Conference (IIPWM)*, pages 369–378.

U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. 2012. Topical clustering of search results. In *5th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 223–232.

O. Zamir and O. Etzioni. 1998. Web document clustering: A feasibility demonstration. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 46–54.

H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma. 2004. Learning to cluster web search results. In *27th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 210–217.