Post-Retrieval Clustering Using Third-Order Similarity Measures

José 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

Guillaume Cleuziou
University of Orléans
LIFO
F-45067 Orléans, France
cleuziou@univ-orleans.fr

Abstract

Post-retrieval clustering is the task of clustering Web search results. Within this context, we propose a new methodology that adapts the classical K-means algorithm to a third-order similarity measure initially developed for NLP tasks. Results obtained with the definition of a new stopping criterion over the ODP-239 and the MORESQUE gold standard datasets evidence that our proposal outperforms all reported text-based approaches.

1 Introduction

Post-retrieval clustering (PRC), also known as search results clustering or ephemeral clustering, is the task of clustering Web search results. For a given query, the retrieved Web snippets are automatically clustered and presented to the user with meaningful labels in order to minimize the information search process. This technique can be particularly useful for polysemous queries but it is hard to implement efficiently and effectively (Carpineto et al., 2009). Indeed, as opposed to classical text clustering, PRC must deal with small collections of short text fragments (Web snippets) and be processed in run-time.

As a consequence, most of the successful methodologies follow a monothetic approach (Zamir and Etzioni, 1998; Ferragina and Gulli, 2008; Carpineto and Romano, 2010; Navigli and Crisafulli, 2010; Scaiella et al., 2012). The underlying idea is to discover the most discriminant topical words in the collection and group together Web snippets containing these relevant terms. On the other hand, the polythetic approach which main idea is to represent Web snippets as word feature vectors has received less attention, the only relevant work being (Osinski and Weiss, 2005). The main reasons for this situation are that (1) word feature vectors are hard to define in small collections of short text fragments (Timonen, 2013), (2) existing second-order similarity measures such as the cosine are unadapted to capture the semantic similarity between small texts, (3) Latent Semantic Analysis has evidenced inconclusive results (Osinski and Weiss, 2005) and (4) the labeling process is a surprisingly hard extra task (Carpineto et al., 2009).

This paper is motivated by the fact that the polythetic approach should lead to improved results if correctly applied to small collections of short text fragments. For that purpose, we propose a new methodology that adapts the classical K-means algorithm to a third-order similarity measure initially developed for Topic Segmentation (Dias et al., 2007). Moreover, the adapted K-means algorithm allows to label each cluster directly from its centroids thus avoiding the abovementioned extra task. Finally, the evolution of the objective function of the adapted K-means is modeled to automatically define the "best" number of clusters.

Finally, we propose different experiments over the ODP-239 (Carpineto and Romano, 2010) and MORESQUE (Navigli and Crisafulli, 2010) datasets against the most competitive text-based PRC algorithms: STC (Zamir and Etzioni, 1998), LINGO (Osinski and Weiss, 2005), OPTIMSRC (Carpineto and Romano, 2010) and the classical bisecting incremental K-means (which may be seen as a baseline for the polythetic paradigm)[1]. A new evaluation measure called the b-cubed F-measure (F_{b^3}) and defined in (Amigó et al., 2009) is then calculated to evaluate both cluster homogeneity and completeness. Results evidence that our proposal outperforms all state-of-the-art approaches with a maximum $F_{b^3} = 0.452$ for ODP-239 and $F_{b^3} = 0.490$ for MORESQUE.

[1] The TOPICAL algorithm proposed by (Scaiella et al., 2012) is a knowledge-driven methodology based on Wikipedia.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 153–158,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

2 Polythetic Post-Retrieval Clustering

The K-means is a geometric clustering algorithm (Lloyd, 1982). Given a set of n data points, the algorithm uses a local search approach to partition the points into K clusters. A set of K initial cluster centers is chosen. Each point is then assigned to the center closest to it and the centers are recomputed as centers of mass of their assigned points. The process is repeated until convergence. To assure convergence, an objective function Q is defined which decreases at each processing step. The classical objective function is defined in Equation (1) where π_k is a cluster labeled k, $x_i \in \pi_k$ is an object in the cluster, m_{π_k} is the centroid of the cluster π_k and $E(.,.)$ is the Euclidean distance.

$$Q = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} E(x_i, m_{\pi_k})^2. \tag{1}$$

Within the context of PRC, the K-means algorithm needs to be adapted to integrate third-order similarity measures (Mihalcea et al., 2006; Dias et al., 2007). Third-order similarity measures, also called weighted second-order similarity measures, do not rely on exact matches of word features as classical second-order similarity measures (e.g. the cosine metric), but rather evaluate similarity based on related matches. In this paper, we propose to use the third-order similarity measure called InfoSimba introduced in (Dias et al., 2007) for Topic Segmentation and implement its simplified version S_s^3 in Equation 2.

$$S_s^3(X_i, X_j) = \frac{1}{p^2} \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik} * X_{jl} * S(W_{ik}, W_{jl}). \tag{2}$$

Given two Web snippets X_i and X_j, their similarity is evaluated by the similarity of its constituents based on any symmetric similarity measure $S(.,.)$ where W_{ik} (resp. W_{jl}) corresponds to the word at the k^{th} (resp. l^{th}) position in the vector X_i (resp. X_j) and X_{ik} (resp. X_{jl}) is the weight of word W_{ik} (resp. W_{jl}) in the set of retrieved Web snippets. A direct consequence of the change in similarity measure is the definition of a new objective function $Q_{S_s^3}$ to ensure convergence. This function is defined in Equation (3) and must be maximized[2].

$$Q_{S_s^3} = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} S_s^3(x_i, m_{\pi_k}). \tag{3}$$

A cluster centroid m_{π_k} is defined by a vector of p words $(w_1^{\pi_k}, \ldots, w_p^{\pi_k})$. As a consequence, each cluster centroid must be instantiated in such a way that $Q_{S_s^3}$ increases at each step of the clustering process. The choice of the best p words representing each cluster is a way of assuring convergence. For that purpose, we define a procedure which consists in selecting the best p words from the global vocabulary V in such a way that $Q_{S_s^3}$ increases. The global vocabulary is the set of all words which appear in any context vector.

So, for each word $w \in V$ and any symmetric similarity measure $S(.,.)$, its interestingness $\lambda^k(w)$ is computed as regards to cluster π_k. This operation is defined in Equation (4) where $s_i \in \pi_k$ is any Web snippet from cluster π_k. Finally, the p words with higher $\lambda^k(w)$ are selected to construct the cluster centroid. In such a way, we can easily prove that $Q_{S_s^3}$ is maximized. Note that a word which is not part of cluster π_k may be part of the centroid m_{π_k}.

$$\lambda^k(w) = \frac{1}{p} \sum_{s_i \in \pi_k} \sum_{w_q^i \in s_i} S(w_q^i, w). \tag{4}$$

Finally, we propose to rely on a modified version of the K-means algorithm called Global K-means (Likasa et al., 2003), which has proved to lead to improved results. To solve a clustering problem with M clusters, all intermediate problems with $1, 2, ..., M-1$ clusters are sequentially solved. The underlying idea is that an optimal solution for a clustering problem with M clusters can be obtained using a series of local searches using the K-means algorithm. At each local search, the $M-1$ cluster centers are always initially placed at their optimal positions corresponding to the clustering problem with $M-1$ clusters. The remaining M^{th} cluster center is initially placed at several positions within the data space. In addition to effectiveness, the method is deterministic and does not depend on any initial conditions or empirically adjustable parameters. Moreover, its adaptation to PRC is straightforward.

3 Stopping Criterion

Once clustering has been processed, selecting the best number of clusters still remains to be decided.

[2] A maximization process can easily be transformed into a minimization one

For that purpose, numerous procedures have been proposed (Milligan and Cooper, 1985). However, none of the listed methods were effective or adaptable to our specific problem. So, we proposed a procedure based on the definition of a rational function which models the quality criterion $Q_{S_s^3}$. To better understand the behaviour of $Q_{S_s^3}$ at each step of the adapted GK-means algorithm, we present its values for $K = 10$ in Figure 1.

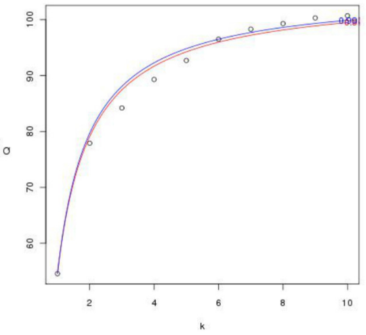

Figure 1: $Q_{S_s^3}$ and its modelisation.

$Q_{S_s^3}$ can be modelled as in Equation (5) which converges to a limit α when K increases and starts from $Q^1_{S_s^3}$ (i.e. $Q_{S_s^3}$ at $K = 1$). The underlying idea is that the best number of clusters is given by the β value which maximizes the difference with the average β^{mean}. So, α, β and γ need to be expressed independently of unknown variables.

$$\forall K, f(K) = \alpha - \frac{\gamma}{K^{\beta}}. \tag{5}$$

As α can theoretically or operationally be defined and it can easily be proved that $\gamma = \alpha - Q^1_{S_s^3}$, β needs to be defined based on γ or α. This can also be easily proved and the given result is expressed in Equation (6).

$$\beta = \frac{log(\alpha - Q^1_{S_s^3}) - log(\alpha - Q^K_{S_s^3})}{log(K)}. \tag{6}$$

Now, the value of α which best approximates the limit of the rational function must be defined. For that purpose, we computed its maximum theoretical and experimental values as well as its approximated maximum experimental value based on the δ^2-Aitken (Aitken, 1926) procedure to accelerate convergence as explained in (Kuroda et al., 2008). Best results were obtained with the maximum experimental value which is defined as building the cluster centroid m_{π_k} for each Web snippet individually. Finally, the best number of clusters is defined as in Algorithm (1) and each one receives its label based on the p words with greater interestingness of its centroid m_{π_k}.

Algorithm 1 The best K selection procedure.

```
1. Calculate $\beta^K$ for each $K$
2. Evaluate the mean of all $\beta^K$ i.e. $\beta^{mean}$
3. Select $\beta^K$ which maximizes $\beta^K - \beta^{mean}$
4. Return $K$ as the best number of partitions
```

This situation is illustrated in Figure (1) where the red line corresponds to the rational functional for β^{mean} and the blue line models the best β value (i.e. the one which maximizes the difference with β^{mean}). In this case, the best number would correspond to β^6 and as a consequence, the best number of clusters would be 6. In order to illustrate the soundness of the procedure, we present the different values for β at each K iteration and the differences between consecutive values of β at each iteration in Figure 2. We clearly see that the highest inclination of the curve is between cluster 5 and 6 which also corresponds to the highest difference between two consecutive values of β.

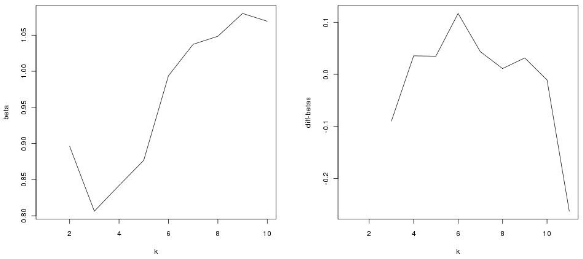

Figure 2: Values of β (on the left) and differences between consecutive values of β (on the right).

4 Evaluation

Evaluating PRC systems is a difficult task as stated in (Carpineto et al., 2009). Indeed, a successful PRC system must evidence high quality level clustering. Ideally, each query subtopic should be represented by a unique cluster containing all the relevant Web pages inside. However, this task is far from being achievable. As such, this constraint is reformulated as follows: the task of PRC systems is to provide complete topical cluster coverage of a given query, while avoiding excessive

F_{b^3}			K									Stop Criterion	
			2	3	4	5	6	7	8	9	10	F_{b^3}	Avg. K
SCP	p	2	0.387	0.396	0.398	0.396	0.391	0.386	0.382	0.378	0.374	0.395	4.799
		3	0.400	0.411	0.412	0.409	0.406	0.400	0.397	0.391	0.388	0.411	4.690
		4	0.405	0.416	0.423	0.425	0.423	0.420	0.416	0.414	0.411	0.441	4.766
		5	0.408	0.422	**0.431**	**0.431**	0.429	0.429	0.423	0.422	0.421	**0.452**	4.778
PMI	p	2	0.391	0.399	0.397	0.393	0.388	0.383	0.377	0.373	0.366	0.393	4.778
		3	0.408	0.418	0.422	0.418	0.414	0.410	0.405	0.398	0.392	0.416	4.879
		4	0.420	0.434	0.439	0.439	0.435	0.430	0.425	0.420	0.412	0.436	4.874
		5	0.423	0.444	**0.451**	**0.451**	**0.451**	0.445	0.441	0.434	0.429	**0.450**	4.778

Table 1: F_{b^3} for SCP and PMI for the global search and the stopping criterion for the ODP-239 dataset.

		Adapated GK-means								STC	LINGO	BIK	OPTIMSRC
		SCP				PMI							
		p				p							
		2	3	4	5	2	3	4	5				
ODP-239	F_1	0.312	0.341	0.352	0.366	0.332	0.358	0.378	**0.390**	0.324	0.273	0.200	0.313
	F_2	0.363	0.393	0.404	0.416	0.363	0.395	0.421	**0.435**	0.319	0.167	0.173	0.341
	F_5	0.411	0.441	0.453	0.462	0.390	0.430	0.459	**0.476**	0.322	0.153	0.165	0.380
	F_{b^3}	0.395	0.411	0.441	**0.452**	0.393	0.416	0.436	0,450	0.403	0.346	0.307	N/A
MORESQUE	F_1	0.627	0.649	**0.665**	0.664	0.615	0.551	0.543	0.571	0.455	0.326	0.317	N/A
	F_2	0.685	0.733	0.767	**0.770**	0.644	0.548	0.521	0.551	0.392	0.260	0.269	N/A
	F_5	0.747	0.817	0.865	**0.872**	0.679	0.563	0.519	0.553	0.370	0.237	0.255	N/A
	F_{b^3}	0.482	0.482	0.473	0.464	**0.490**	0.465	0.462	0.485	0.460	0.399	0.315	N/A

Table 2: PRC comparative results for F_β and F_{b^3} over the ODP-239 and MORESQUE datasets.

redundancy of the subtopics in the result list of clusters. So, in order to evaluate our methodology, we propose two different evaluations. First, we want to evidence the quality of the stopping criterion when compared to an exhaustive search over all tunable parameters. Second, we propose a comparative evaluation with existing state-of-the-art algorithms over gold standard datasets and recent clustering evaluation metrics.

4.1 Text Processing

Before the clustering process takes place, Web snippets are represented as word feature vectors. In order to define the set of word features, the Web service proposed in (Machado et al., 2009) is used[3]. In particular, it assigns a relevance score to any token present in the set of retrieved Web snippets based on the analysis of left and right token contexts. A specific threshold is then applied to withdraw irrelevant tokens and the remaining ones form the vocabulary V. Then, each Web snippet is represented by the set of its p most relevant tokens in the sense of the $W(.)$ value proposed in (Machado et al., 2009). Note that within the proposed Web service, multiword units are also identified. They are exclusively composed of relevant individual tokens and their weight is given by the arithmetic mean of their constituents scores.

4.2 Intrinsic Evaluation

The first set of experiments focuses on understanding the behaviour of our methodology within a greedy search strategy for different tunable parameters defined as a tuple $< p, K, S(W_{ik}, W_{jl}) >$. In particular, p is the size of the word feature vectors representing both Web snippets and centroids ($p = 2..5$), K is the number of clusters to be found ($K = 2..10$) and $S(W_{ik}, W_{jl})$ is the collocation measure integrated in the InfoSimba similarity measure. In these experiments, two association measures which are known to have different behaviours (Pecina and Schlesinger, 2006) are tested. We implement the Symmetric Conditional Probability (Silva et al., 1999) in Equation (7) which tends to give more credits to frequent associations and the Pointwise Mutual Information (Church and Hanks, 1990) in Equation (8) which over-estimates infrequent associations. Then, best $< p, K, S(W_{ik}, W_{jl}) >$ configurations are compared to our stopping criterion.

$$SCP(W_{ik}, W_{jl}) = \frac{P(W_{ik}, W_{jl})^2}{P(W_{ik}) \times P(W_{jl})}. \quad (7)$$

$$PMI(W_{ik}, W_{jl}) = log_2 \frac{P(W_{ik}, W_{jl})}{P(W_{ik}) \times P(W_{jl})}. \quad (8)$$

In order to perform this task, we evaluate performance based on the F_{b^3} measure defined in (Amigó et al., 2009) over the ODP-239 gold standard dataset proposed in (Carpineto and Romano,

[3] Access to this Web service is available upon request.

2010). In particular, (Amigó et al., 2009) indicate that common metrics such as the F_β-measure are good to assign higher scores to clusters with high homogeneity, but fail to evaluate cluster completeness. First results are provided in Table 1 and evidence that the best configurations for different $< p, K, S(W_{ik}, W_{jl}) >$ tuples are obtained for high values of p, K ranging from 4 to 6 clusters and PMI steadily improving over SCP. However, such a fuzzy configuration is not satisfactory. As such, we proposed a new stopping criterion which evidences coherent results as it (1) does not depend on the used association measure ($F_{b^3}^{SCP} = 0.452$ and $F_{b^3}^{PMI} = 0.450$), (2) discovers similar numbers of clusters independently of the length of the p-context vector and (3) increases performance with high values of p.

4.3 Comparative Evaluation

The second evaluation aims to compare our methodology to current state-of-the-art text-based PRC algorithms. We propose comparative experiments over two gold standard datasets (ODP-239 (Carpineto and Romano, 2010) and MORESQUE (Di Marco and Navigli, 2013)) for STC (Zamir and Etzioni, 1998), LINGO (Osinski and Weiss, 2005), OPTIMSRC (Carpineto and Romano, 2010) and the Bisecting Incremental K-means (BIK) which may be seen as a baseline for the polythetic paradigm. A brief description of each PRC algorithm is given as follows.

STC: (Zamir and Etzioni, 1998) defined the Suffix Tree Clustering algorithm which is still a difficult standard to beat in the field. In particular, they propose a monothetic clustering technique which merges base clusters with high string overlap. Indeed, instead of using the classical Vector Space Model (VSM) representation, they propose to represent Web snippets as compact tries.

LINGO: (Osinski and Weiss, 2005) proposed a polythetic solution called LINGO which takes into account the string representation proposed by (Zamir and Etzioni, 1998). They first extract frequent phrases based on suffix-arrays. Then, they reduce the term-document matrix (defined as a VSM) using Single Value Decomposition to discover latent structures. Finally, they match group descriptions with the extracted topics and assign relevant documents to them.

OPTIMSRC: (Carpineto and Romano, 2010) showed that the characteristics of the outputs returned by PRC algorithms suggest the adoption of a meta clustering approach. As such, they introduce a novel criterion to measure the concordance of two partitions of objects into different clusters based on the information content associated to the series of decisions made by the partitions on single pairs of objects. Then, the meta clustering phase is casted to an optimization problem of the concordance between the clustering combination and the given set of clusterings.

With respect to implementation, we used the Carrot2 APIs[4] which are freely available for STC, LINGO and the classical BIK. It is worth noticing that all implementations in Carrot2 are tuned to extract exactly 10 clusters. For OPTIMSRC, we reproduced the results presented in the paper of (Carpineto and Romano, 2010) as no implementation is freely available. The results are illustrated in Table 2 including both F_β-measure and F_{b^3}. They evidence clear improvements of our methodology when compared to state-of-the-art text-based PRC algorithms, over both datasets and all evaluation metrics. But more important, even when the p-context vector is small ($p = 3$), the adapted GK-means outperforms all other existing text-based PRC which is particularly important as they need to perform in real-time.

5 Conclusions

In this paper, we proposed a new PRC approach which (1) is based on the adaptation of the K-means algorithm to third-order similarity measures and (2) proposes a coherent stopping criterion. Results evidenced clear improvements over the evaluated state-of-the-art text-based approaches for two gold standard datasets. Moreover, our best F_1-measure over ODP-239 (0.390) approximates the highest ever-reached F_1-measure (0.413) by the TOPICAL knowledge-driven algorithm proposed in (Scaiella et al., 2012)[5]. These results are promising and in future works, we propose to define new knowledge-based third-order similarity measures based on studies in entity-linking (Ferragina and Scaiella, 2010).

[4] http://search.carrot2.org/stable/search [Last access: 15/05/2013].

[5] Notice that the authors only propose the F_1-measure although different results can be obtained for different F_β-measures and F_{b^3} as evidenced in Table 2.

References

A.C. Aitken. 1926. On bernoulli's numerical solution of algebraic equations. *Research Society Edinburgh*, 46:289–305.

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. 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, S. Osinski, G. Romano, and D. Weiss. 2009. A survey of web clustering engines. *ACM Computer Survey*, 41(3):1–38.

K. Church and P. Hanks. 1990. Word association norms mutual information and lexicography. *Computational Linguistics*, 16(1):23–29.

A. Di Marco and R. Navigli. 2013. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(4):1–43.

G. Dias, E. Alves, and J.G.P. Lopes. 2007. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of 22nd Conference on Artificial Intelligence (AAAI)*, pages 1334–1339.

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

P. Ferragina and U. Scaiella. 2010. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628.

M. Kuroda, M. Sakakihara, and Z. Geng. 2008. Acceleration of the em and ecm algorithms using the aitken δ^2 method for log-linear models with partially classified data. *Statistics & Probability Letters*, 78(15):2332–2338.

A. Likasa, Vlassis. N., and J. Verbeek. 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36:451–461.

S.P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

D. Machado, T. Barbosa, S. Pais, B. Martins, and G. Dias. 2009. Universal mobile information retrieval. In *Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction (HCI)*, pages 345–354.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 775–780.

G.W. Milligan and M.C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

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 and D. Weiss. 2005. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54.

P. Pecina and P. Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, pages 651–658.

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

J. Silva, G. Dias, S. Guilloré, and J.G.P. Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of 9th Portuguese Conference in Artificial Intelligence (EPIA)*, pages 113–132.

M. Timonen. 2013. *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. Ph.D. thesis, University of Helsinki, Finland.

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.