Search Results Clustering without External Resources

Chris Staff, Joel Azzopardi, Colin Layfield, and Dan Mercieca
Faculty of Information and Communication Technology
University of Malta
Msida MSD 2080, Malta
Email: {chris.staff,joel.azzopardi,colin.layfield,daniel.mercieca.12}@um.edu.mt

***Abstract*—Our unsupervised Search Results Clustering (SRC) system partitions into clusters the top-n results returned by a search engine. We present the results of experiments with our SRC system that performs incremental clustering on document titles and snippets only and does not use external resources, yet which outperforms the best performers to date on the SemEval-2013 Task 11 gold standard. We include Latent Semantic Analysis (LSA) as an optional step, using the snippets themselves as the background corpus. We demonstrate that better results are achieved by leaving the query terms out of the clustering process, and that currently, the version without LSA outperforms the version with LSA.**

***Keywords*—SemEval-2013 Task 11; K-Means; Unsupervised Clustering; Generalized Dunn's Index; Latent Semantic Analysis; Word Sense Induction;**

I. Introduction

Search Results Clustering (SRC) changes how search engine results lists are presented to users so they can more easily identify subjectively relevant results.

When presented with a list of search results, users can identify some possibly relevant documents by their titles and snippets. However, results lists also contain non-relevant results as queries may be underspecified or ambiguous [1], [2]. Search engines diversify results to increase the chances of providing some relevant results [3]. However, it is likely that some relevant results i) may not have been retrieved because they do not contain any query terms; ii) may not be seen by the user because they are ranked too low in the results list; and, iii) may be overlooked by the user because they are surrounded by largely non-relevant results. By partitioning the results into those that are subjectively relevant and those that are not we can alleviate some of these problems [4].

Our clustering algorithm, No-K-Means, is incremental and unsupervised. We process the document title and snippet only, without using external language resources to guide the clustering process. We experiment with and without Latent Semantic Analysis (LSA) [5] to identify snippets in the results list that may be semantically similar but that do not necessarily contain similar terms. We also experiment with including and excluding the query terms when determining cluster membership. We generate non-overlapping clusters. No-K-Means does not need to be given the number of clusters to create beforehand. We use Generalized Dunn's Index [6] to perform an internal evaluation of generated clusters to discover the best cluster configuration.

We evaluate the quality of clusters generated by our approach using the SemEval-2013 Task 11 gold standard collection [7]. The Word Sense Induction Evaluator supplied with the collection measures the F1, Rand Index (RI), Adjusted Rand Index (ARI), and Jaccard Index (JI) scores for the generated clusters compared to the gold standard. Our results show that omitting the query terms significantly improves the quality of the clusters generated. Our approach also outperforms the best performers in the SemEval-2013 Task 11 Workshop and others who have subsequently experimented with the same collection.

We discuss relevant literature in Sect. II, and our approach in Sect. III. In Sect. IV we describe the evaluation and results of the clustering process. We give our conclusions and plans for future work in Sect. V.

II. Literature Review

Carpineto, et al. provide a relatively recent, and very detailed, survey of the state-of-the-art in Web clustering, including Search Results Clustering (SRC) [4]. In this paper, we focus on different general approaches to Search Results Clustering (Sect. II-A) and approaches taken by participants in the SemEval-2013 Task 11 workshop (Sect. II-B).

A. Search Results Clustering

In SRC in general, the documents to cluster are the top-n ranked documents in a results list returned by a search engine following a user query. Queries can be short and query terms may be polysemous or ambiguous so results may not all be about the user's intended meaning. Clustering can be used to group together documents that are relevant to the different word senses of the query terms.

The evidence on which clustering is performed can be snippets of text returned by a search engine [8], [9], [10], or the full-text of documents [11], [12]. Singh, et al. perform clustering based on the first 200 words in each document [13]. Often, to improve the clustering process, external resources such as DBPedia [14], Wikipedia [15], [12], query logs [16], [17], [18], web sites [17], the Open Data Project (ODP) [3], [13], and TAGME [12], are used. Cluster *centroids* or descriptions can be identified to represent documents that should be classified into each cluster. Candidate centroids can be extracted or generated from common phrases that occur in the collection [19], [8], [10]. Alternatively, K-Means [20] or one of its variants can be used to allocate a random seed document to each of K clusters, classify the remaining documents on the basis of the seeds, and modify the cluster membership over subsequent passes through the collection [8], [21]. As [19] point out, "there is no optimal predefined K fit for all queries" to choose the right number of clusters in advance.

Once cluster centroids have been chosen, a classifier determines cluster membership. The choice of classifier will have a bearing on the document preprocessing necessary to facilitate the classification. Search engines typically retrieve documents that contain the terms expressed in a user query (objective relevance) but not all of the documents are likely to be subjectively relevant to the user due to term ambiguity. Similarly, cluster quality will suffer if the documents in a cluster share the same terms at a syntactic level but are not semantically related. Also, semantically related documents may be classified into different clusters if the classifier uses only term similarity at a syntactic level, which means that documents relevant to a user could be spread across many different clusters.

Latent Dirichlet Allocation (LDA) can discover latent topics described by texts, even when the document terms are syntactically different. However, usually a large document collection is required to discover the underlying relationships [15]. Latent Semantic Analysis (LSA) [11], [13], [5] can discover semantic relatedness between different terms in a collection based on their co-occurrence with common terms. Search Result Clustering approaches based on LSA typically utilise a background corpus, larger than and different from the documents being analysed, to independently discover semantic relatedness between terms. In LDA, the documents in the search results list are mapped onto a topic hierarchy to yield clusters. In LSA, the semantically enriched document representations are used, potentially clustering together related documents that do not have terms in common. Combining LSA with TFIDF (Term Frequency × Inverse Document Frequency) can lead to improvements in cluster quality [13]. TFIDF is a typical way of deriving term weights in the standard "bag-of-words" approach. [13] chooses, for each cluster, a label that consists of the three terms that are most common in the cluster, but they use the Open Directory Project as an external resource to guide the clustering process.

Clusters can be non-overlapping, meaning that a document is assigned to one cluster only (i.e., the one to which it is most similar) [21], or overlapping, meaning that a document is assigned to all clusters to which it is similar enough (i.e., document-cluster similarity exceeds some threshold) [12]. Finally, the clusters may be refined by merging clusters that have a large number of documents in common [21], [22]; splitting clusters [22]; or merging clusters each containing a single document (singletons) into an 'Others' cluster [8].

B. The SemEval-2013 Task 11 Workshop

Navigli et al. [7] organised the SemEval-2013 Task 11 Workshop 'Word Sense Induction and Disambiguation within an End-User Application'. The chosen application was Search Results Clustering. Workshop participants devised different approaches to attempt to cluster results by word sense. As SRC is difficult to evaluate, Navigli et al. constructed a gold standard collection comprising 100 ambiguous topics which were submitted to the Google search engine as queries to retrieve 64 results each. The results were manually organised into 'subtopics' or clusters. A result comprises a document URL, title, and snippet. The organisers also provided an automatic evaluator, described in [7], to measure cluster quality and accuracy using F1, Rand Index (RI), Adjusted Rand Index (ARI), and Jaccard Index (JI).

With the exception of SATTY [23], the Workshop participants adopted cluster creation approaches that did not require them to know in advance the number of clusters to create. SATTY, however, attempted to identify the ten most diverse snippets in each results list which were then used as cluster centroids. The University of Melbourne [24], UKP-WSI [25], and Duluth [26] used background corpora, derived and processed in different ways, to discover contextually related terms commonly co-located with the different possible senses of the terms in the queries to classify snippets according to their possible word sense. The University of Melbourne used Hierarchical Dirichlet Processing (HDP) to discover topics in a collection based on the English Wikipedia dump taken in November 2011, and their approaches performed best in the Workshop. The University of Melbourne submitted two systems: HDP-Clusters-Lemma lemmatises terms whereas in the HDP-Clusters-NoLemma approach, terms are not lemmatised. Duluth applied LSA to three different background collections (i. the SemEval results snippets for the query only; ii. the result snippets of all 100 SemEval queries; and, iii. a subset of the Gigaword collection). Their best performance was achieved using the first background collection. UKP-WSI used an English Wikipedia dump and ukWaC as different background corpora to derive and utilise term co-occurrence statistics to determine word sense. Additionally, they used the entire document text, when available, rather than just the snippet.

SenseSearcher (SnS) [9] did not participate in the workshop but used the SemEval collection to evaluate their Word Sense Induction algorithm. They discover a hierarchy of word senses from a raw text corpus, composed of the snippets in the results list for each SemEval query processed to do Part-of-Speech tagging and to identify proper nouns. Word senses of a term are disambiguated by examining the 'context' surrounding a term, where the context is the co-occurring terms in snippets.

III. The No-K-Means Algorithm

Our No-K-Means clustering algorithm clusters the top-n results retrieved by a search engine. We have two versions of the algorithm, one which uses LSA (withLSA) and the other which does not (noLSA). In both versions, each document title and snippet pair in the results list is processed to stem terms using the Porter Stemmer [27] and stop words are removed using the Onix Text Retrieval Toolkit's Stop Word List 1[1]. The query terms are also removed, and the result URLs are ignored. The remaining stems are then used to create a term-by-document matrix. Rather than using TFIDF to calculate term weights, we use raw term frequency (TF) only. TFIDF is normally used to dampen the effect of terms that occur frequently in a collection. TFIDF is practical in massive collections when the overwhelming majority of documents are not relevant to a user's query. When a user includes as a query

[1] Available from http://www.lextek.com/manuals/onix/stopwords1.html.

term a term that is highly pervasive in the collection, the term itself is probably not a good discriminator between relevant and non-relevant documents, nor is it helpful because the majority of documents in the collection would be retrieved. In our case, and in the context of SRC, the 'collection' is the top-n snippets that have been retrieved as part of a results set following a user query. The snippets that are retrieved are a small subset of the entire collection (both in terms of the entire results set, and in terms of the entire document space). We remove the query terms from the representations of the snippets as the query terms do not help with partitioning the results set into clusters. We show (in Sect. IV-A) that retaining query terms severely degrades cluster quality. Consequently, we use only TF as the term weights in snippets, which also means that we can calculate term weights without needing to process all of the snippets in advance (to work out in how many the term occurs and then derive IDF). Currently, we do not distinguish between terms occurring in the title and in the snippet when calculating the term weight.

In the 'withLSA' version, we perform singular-value decomposition on the term-by-document matrix using an adaptation of the code from the Numerical Recipes SVD implementation available at http://www.nr.com/webnotes/nr3web2.pdf. The output from the LSA engine is a k-rank matrix semantic space approximation of the original matrix. Removing the query terms from the document titles and snippets eliminates the possibility of over-associating terms across documents due to the query terms that are pervasive and that do not help with the clustering process. Following the LSA step, each document is transformed into a lower dimensional representation, when compared to the original matrix, that may expose the 'latent semantics' of the document when compared with others in the same semantic space to reflect an affinity between related terms. The resulting document vectors are then processed by our clustering algorithm. The remainder of the algorithm is performed by both the 'withLSA' and 'noLSA' versions.

Our No-K-Means clustering algorithm does not require the target number of clusters to be known in advance (hence the name No-K). The first document (title, snippet pair) in the results list is put into its own cluster and the document vector is the initial cluster centroid. The vector of the next document in the results list is compared to the existing cluster centroids using the standard cosine similarity measure and it is placed into the most similar cluster, given some threshold. The cluster centroid is recomputed (a simple averaging of term weights). If the document is not similar enough to any existing cluster a new cluster is created for it. Once all the documents have been processed, the singleton clusters (clusters containing just one document) are merged into an 'Others' cluster. Thus, No-K-Means produces non-overlapping clusters.

No-K-Means is derived from an earlier version of a clustering algorithm we developed for the online incremental clustering of news reports (JNews [28]). The significant differences between the JNews algorithm and No-K-Means are: JNews was designed to process a stream of global news reports and generate clusters according to news event. It processed the full-text of the reports, and weighted terms using TFIDF where IDF was calculated based on the entire document collection processed to date. In No-K-Means, we process short text fragments (snippets and titles) that are related to the user's original query terms (though not necessarily to the user's intended meaning of those terms), we weight terms using TF only, and we remove the query terms from the representations. Additionally, No-K-Means has an optional LSA step that is not present in the JNews algorithm.

IV. Evaluation and Results

Cluster quality is typically evaluated by comparing the generated clusters to a gold standard [8], [4], [22], [14], [18], [12], [21]. The latest gold standards used in the evaluation of Search Results Clustering are SemEval-2013 Task 11 [7], [18], [22] and WEBSRC401 [18]. The ODP-239 dataset [8], [21], [22], [18], [12], its predecessor (AMBIENT [8], [4], [22], [14], [18], [12]) and a related dataset (DMOZ-50 [22]) have also been used. The latter datasets are derived from the Open Directory Project, a document collection hierarchically organised by topics. In SemEval the topics are organised by specific query (e.g., SemEval's 'marble->Marble sculpture, the art of creating three-dimensional forms from marble'), but in ODP-239, AMBIENT, and DMOZ-50, the datasets are organised into a canonical hierarchy of generic topics (e.g., 'Arts->Architecture->Education'). We consider the SemEval collection to contain 'topics' that are more typical of user queries submitted to a search engine, also observing that [12] claims that it is difficult to use OPD-239 "because sub-topics are very similar to each other and textual fragments are very short". We consider AMBIENT and DMOZ-50 to be equally difficult to use for Search Results Clustering. WEBSRC401 has a format similar to SemEval's but has overlapping clusters, unlike SemEval. Our algorithm yields non-overlapping clusters so we cannot evaluate our approach on WEBSRC401.

As introduced in Sect. II-B, SemEval contains 100 'topics' of between one and four keywords each extracted from Wikipedia's list of ambiguous topics. The collection was created by submitting the topics as queries to Google's search engine in 2012, collecting the snippets of the top 64 results (6400 snippets in total), organising them into sub-topics, and annotating them with subtopic relevance judgements [14]. On average, there are 7.69 subtopics (clusters) per topic (query). The publicly available dataset[2] contains an evaluator that measures the quality of the clusters generated using F1, Jaccard Index, Rand Index, and Adjusted Rand Index [7].

A. Our Cluster Quality Experiments and Results

We experimented with different configurations of No-K-Means withLSA (varying the number of LSA dimensions, k, in steps of 5) and noLSA, including and excluding the query terms, and the similarity threshold, *simThres*:

- withLSA, noQT (no query terms), $5 \leq k \leq 60$, 0.01 $\leq$ *simThres* $\leq$ 0.9.
- withLSA, withQT (with query terms), $5 \leq k \leq 60$, 0.01 $\leq$ *simThres* $\leq$ 0.9.
- noLSA, noQT, 0.01 $\leq$ *simThres* $\leq$ 0.9.
- noLSA, withQT, 0.01 $\leq$ *simThres* $\leq$ 0.9.

In all cases, the similarity threshold (*simThres*) was incremented by 0.01 between the values 0.01 and 0.09, and then

[2] http://www.cs.york.ac.uk/semeval-2013/task11/.

TABLE I. BEST PERFORMERS ON THE SEMEVAL-2013 TASK 11 DATASET

	F1	**RI**	**ARI**	**JI**	**Ave. # clusters**	**Ave. clus. size**
hdp-clusters-lemma	68.30	65.22	21.31	33.02	6.63	11.07
hdp-clusters-nolemma	68.03	64.86	21.49	33.75	6.54	11.68
SnS	**70.16**	**65.84**	**22.19**	**34.26**	8.82	8.46
SemEval singletons	100.00	60.09	0.00	0.00	64.00	1.00
SemEval all-in-one	54.42	39.90	0.00	39.90	1.00	64.00
Gold Standard					**7.69**	**11.56**

TABLE II. OUR RESULTS WITH NO-K-MEANS - NOLSA, WITHOUT AND WITH QUERY TERMS (QT). SCORES THAT BEAT THE BEST PERFORMERS TO DATE ARE IN BOLD.

QT	***SimThres***	**F1**	**RI**	**ARI**	**JI**	**Ave. # clusters**	**Ave. clus. size**
	GDI_Varied	**71.86**	**68.59**	**26.67**	**35.47**	7.96	9.52
	GDI_Fixed	**72.27**	**68.83**	**27.67**	**35.66**	**7.89**	8.90
	0.01	64.75	62.11	19.43	**36.87**	5.01	14.72
✓	GDI_Varied	56.70	43.46	3.57	38.49	2.48	27.63
✓	GDI_Fixed	54.61	40.23	0.00	39.85	1.17	58.77
✓	0.01	54.58	40.22	0.00	39.92	1.11	60.69

TABLE III. OUR RESULTS WITH NO-K-MEANS - WITHLSA, WITHOUT AND WITH QUERY TERMS (QT).

QT	***Mode***	**F1**	**RI**	**ARI**	**JI**	**Ave. # clusters**	**Ave. clus. size**
	GDI_Varied	64.33	60.63	19.71	38.57	3.90	21.07
	GDI_Fixed	67.92	62.74	18.12	27.44	5.69	11.41
✓	GDI_Varied	54.91	40.68	0.01	39.73	1.27	55.57
✓	GDI_Fixed	54.50	40.18	0.00	39.85	1.08	61.44

in steps of 0.1 to 0.9. *simThres* is the minimum similarity threshold to classify a document into a cluster. We use Bezdek and Pal's Generalized Dunn's Index [29] to objectively find the best values for *simThres* (in withLSA and noLSA) and *k* (in withLSA only). Generalized Dunn's Index (GDI) yields a value based on how 'well behaved' the clusters and cluster members are, measured as a function of between-clusters and within-clusters distances. We calculate the GDI for a set of m clusters $C_1, C_2, \ldots, C_m$ (GDI_m) using Equation 1.

$$GDI_m = \frac{\min\limits_{1 \ll i < j \ll m} \delta(C_i, C_j)}{\max\limits_{1 \ll k \ll m} \Delta_k}$$

where:

$$\Delta_i = 2 \times \frac{\sum\limits_{x \in C_i} \delta(\mu_i, x)}{|C_i|} \qquad \mu_i = \frac{\sum\limits_{x \in C_i} x}{|C_i|}$$

$$\delta(C_i, C_j) = \delta(\mu_i, \mu_j) \tag{1}$$

We use GDI in two ways: i) fixing *simThres* and *k* for all queries in the test collection, and iteratively running all combinations (GDI_Fixed); and, ii) varying *simThres* and *k* per query (GDI_Varied). In GDI_Fixed, we use GDI to identify the *simThres* and *k* that produce the best clusters on average, while GDI_Varied identifies the best *simThres* and *k* per query.

Table I reports the best results obtained at the SemEval-2013 Task 11 Workshop [7], together with the baselines (singletons and all-in-one) provided by the organisers [7]. As discussed in Sect. II-B, the HDP clusters are obtained using Hierarchical Dirichlet Process which is trained using Wikipedia [24], [7]. SenseSearcher (SnS) has subsequently obtained better results, despite being "knowledge-poor": relying only on syntactic parsing of the collection and identifying proper nouns [9].

Our results are presented in Table II (noLSA) and Table III (withLSA). In Table II, both the noLSA configuration objectively chosen based on the GDI value when *simThres* is varied across all queries (GDI_Varied) and when it is fixed (GDI_Fixed) outperform the best performers at the SemEval workshop and SnS on all measures. GDI_Fixed outperforms GDI_Varied on all measures. *simThres* has a value of 0.05 when GDI_m is greatest for noLSA with query terms omitted.

We also show in Table II how results deteriorate when the query terms are included in the snippet representations and clustering process. Indeed, this is clearest in the noLSA version when *simThres* is fixed at 0.01 for all queries. When query terms are included, the No-K-Means performance is in line with the baseline (all-in-one, in Table I). However, when the query terms are omitted, and for the same fixed *simThres*, the noLSA version of No-K-Means creates 5.01 clusters on average, with significantly improved F1, RI, and ARI scores and even these results are competitive compared to those achieved by the SemEval workshop participants. The withLSA version (Table III) does not perform as well as noLSA which suggests that document titles and snippets are too short and on their own are not rich enough to find semantic alternatives ([30] shows some evidence that larger document size is likely to improve the effectiveness of LSA). However, even for withLSA, including query terms in the snippet representations and clustering process harms accuracy.

V. DISCUSSION, CONCLUSION, AND FUTURE WORK

Our No-K-Means approach with noLSA outperforms the best performers to date on the SemEval-2013 Task 11 collection, without using external resources, and without syntactic or semantic processing of the snippets. We perform unsupervised, incremental clustering, without needing to know in advance the number of clusters to generate.

Our results show a significant improvement in cluster quality (measured by F1, RI, and ARI) when query terms are omitted from the snippet representations and clustering process. The withLSA results are not as good as noLSA's although they are competitive compared to the other participants in the SemEval-2013 Task 11 workshop [7]. We will continue to experiment with the withLSA version, to determine if using a different background collection will improve results, and with No-K-Means in general to investigate whether the order in which clusters are created can lead to better results.

REFERENCES

[1] B. Vélez, R. Weiss, M. A. Sheldon, and D. K. Gifford, "Fast and effective query refinement," in *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA, USA*, 1997, pp. 6–15. [Online]. Available: http://doi.acm.org/10.1145/258525.258528

[2] D. C. Anastasiu, B. J. Gao, and D. Buttler, "A framework for personalized and collaborative clustering of search results," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, 2011, pp. 573–582. [Online]. Available: http://doi.acm.org/10.1145/2063576.2063662

[3] D. Kuang, X. Li, and C. X. Ling, "A new search engine integrating hierarchical browsing and keyword search," in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, 2011, pp. 2464–2469. [Online]. Available: http://ijcai.org/papers11/Papers/IJCAI11-410.pdf

[4] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A survey of web clustering engines," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 17:1–17:38, Jul. 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541884

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[6] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, January 1973.

[7] R. Navigli and D. Vannella, "SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 193–201. [Online]. Available: http://www.aclweb.org/anthology/S13-2035

[8] A. Turel and F. Can, "A new approach to search result clustering and labeling," in *Information Retrieval Technology - 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings*, 2011, pp. 283–292. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-25631-8_26

[9] M. Kozłowski and H. Rybiński, "SnS: A novel word sense induction method," in *Rough Sets and Intelligent Systems Paradigms*, ser. Lecture Notes in Computer Science, M. Kryszkiewicz, C. Cornelis, D. Ciucci, J. Medina-Moreno, H. Motoda, and Z. Raś, Eds. Springer International Publishing, 2014, vol. 8537, pp. 258–268. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-08729-0_25

[10] P. Ferragina and A. Gulli, "A personalized search engine based on web-snippet hierarchical clustering," in *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005 - Special interest tracks and posters*, 2005, pp. 801–810. [Online]. Available: http://doi.acm.org/10.1145/1062745.1062760

[11] G. Mecca, S. Raunich, and A. Pappalardo, "A new algorithm for clustering search results," *Data Knowl. Eng.*, vol. 62, no. 3, pp. 504–522, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2006.10.006

[12] U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita, "Topical clustering of search results," in *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, 2012, pp. 223–232. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124324

[13] R. Singh, Y. Hsu, and N. Moon, "Multiple perspective interactive search: a paradigm for exploratory search and information retrieval on the web," *Multimedia Tools Appl.*, vol. 62, no. 2, pp. 507–543, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11042-011-0910-2

[14] M. Schuhmacher and S. P. Ponzetto, "Exploiting DBpedia for web search results clustering," in *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, ser. AKBC '13. New York, NY, USA: ACM, 2013, pp. 91–96. [Online]. Available: http://doi.acm.org/10.1145/2509558.2509574

[15] X. H. Phan, M. L. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, 2008, pp. 91–100. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367510

[16] W. Yih and C. Meek, "Improving similarity measures for short segments of text," in *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, 2007, pp. 1489–1494. [Online]. Available: http://www.aaai.org/Library/AAAI/2007/aaai07-236.php

[17] Z. Dou, S. Hu, K. Chen, R. Song, and J. Wen, "Multi-dimensional search result diversification," in *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, 2011, pp. 475–484. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935897

[18] J. G. Moreno, G. Dias, and G. Cleuziou, "Query log driven web search results clustering," in *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, 2014, pp. 777–786. [Online]. Available: http://doi.acm.org/10.1145/2600428.2609583

[19] Y. Zhang and B. Feng, "A co-occurrence based hierarchical method for clustering web search results," in *2008 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, 9-12 December 2008, Sydney, NSW, Australia, Main Conference Proceedings*, 2008, pp. 407–410. [Online]. Available: http://dx.doi.org/10.1109/WIIAT.2008.35

[20] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[21] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani, "Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution," in *String Processing and Information Retrieval, 13th International Conference, SPIRE 2006, Glasgow, UK, October 11-13, 2006, Proceedings*, 2006, pp. 25–36. [Online]. Available: http://dx.doi.org/10.1007/11880561_3

[22] C. Cobos, H. Muñoz-Collazos, R. Urbano-Muñoz, M. Mendoza, E. L. Guzman, and E. Herrera-Viedma, "Clustering of web search results based on the cuckoo search algorithm and balanced bayesian information criterion," *Inf. Sci.*, vol. 281, pp. 248–264, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2014.05.047

[23] S. Behera, U. Gaikwad, R. Bairi, and G. Ramakrishnan, "SATTY : Word sense induction application in web search clustering," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 207–211. [Online]. Available: http://www.aclweb.org/anthology/S13-2037

[24] J. H. Lau, P. Cook, and T. Baldwin, "Topic modelling-based word sense induction for web snippet clustering," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 217–221. [Online]. Available: http://www.aclweb.org/anthology/S13-2039

[25] H.-P. Zorn and I. Gurevych, "UKP-WSI: UKP lab SemEval-2013 task 11 system description," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 212–216. [Online]. Available: http://www.aclweb.org/anthology/S13-2038

[26] T. Pedersen, "Duluth : Word sense induction applied to web page clustering," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).* Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 202–206. [Online]. Available: http://www.aclweb.org/anthology/S13-2036

[27] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[28] J. Azzopardi and C. Staff, "Incremental clustering of news reports," *Algorithms*, vol. 5, no. 3, pp. 364–378, 2012. [Online]. Available: http://www.mdpi.com/1999-4893/5/3/364

[29] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *Trans. Sys. Man Cyber. Part B*, vol. 28, no. 3, pp. 301–315, Jun. 1998. [Online]. Available: http://dx.doi.org/10.1109/3477.678624

[30] C. Layfield, "With LSA size DOES matter," in *Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation, EMS 2012, Malta, November 14-16, 2012.* IEEE, 2012, pp. 127–131. [Online]. Available: http://dx.doi.org/10.1109/EMS.2012.24