Incorporating Deep Visual Features into Multiobjective based Multi-view Search Results Clustering

Sayantan Mitra
Department of CSE,
IIT Patna,
Bihta, India
sayantaniem@gmail.com

Mohammed Hasanuzzaman
ADAPT Centre,
School of Computing,
Dublin City University,
Dublin, Ireland

Sriparna Saha
Department of CSE,
IIT Patna,
Bihta, India

Andy Way
ADAPT Centre,
School of Computing,
Dublin City University,
Dublin, Ireland

Abstract

Current paper explores the use of multi-view learning for search result clustering. A web-snippet can be represented using multiple views. Apart from textual view cued by both the semantic and syntactic information, a complementary view extracted from images contained in the web-snippets is also utilized in the current framework. A single consensus partitioning is finally obtained after consulting these two individual views by the deployment of a multi-objective based clustering technique. Several objective functions including the values of a cluster quality measure evaluating the goodness of partitionings obtained using different views and an agreement-disagreement index, quantifying the amount of oneness among multiple views in generating partitionings are optimized simultaneously using AMOSA. In order to detect the number of clusters automatically, concepts of variable length solutions and a vast range of permutation operators are introduced in the clustering process. Finally a set of alternative partitionings are obtained on the final Pareto front by the proposed multi-view based multi-objective technique. Experimental results by the proposed approach on several bench-mark test datasets with respect to different performance metrics evidently establish the power of visual and text based views in achieving better search result clustering.

1 Introduction

Web search results clustering (SRC), also known as ephemeral clustering or post-retrieval clustering has garnered much attention in the past few decades for making web browsing easier for users. The key objective of SRC systems is: for a given query it can return some meaningful labeled clusters from a set of web documents (or web snippets) retrieved from a search engine. Recent years have witnessed a large number of attempts in solving this SRC problem (Di Marco and Navigli, 2013; Scaiella et al., 2012). Most of them have developed some clustering algorithms optimizing a single objective criteria (Osinski and Weiss, 2005; Zamir and Etzioni, 1998; Moreno et al., 2013). But a complex data set like set of web-snippets can be clustered into several alternative partitionings. Therefore to detect all possible partitionings containing clusters of different shapes automatically, application of multiobjective optimization (MOO) for solving the problem of clustering becomes prevalent (Maulik et al., 2011). In this context, Acharya et al. (Acharya et al., 2014) have proposed a multi objective optimization based approach for solving SRC problem by extracting both semantic and syntactic information present in web-snippets. Results attained by this approach outperformed the other existing single objective based approaches. Moreover a web-snippet can be represented using different views, for example semantic view, syntactic view. Recently, Wahid et al. (Wahid et al., 2014) have developed a multi-view based MOO clustering technique for search result clustering where multiple views are consulted for developing a consensus partitioning of available web-snippets.

Multimodal approaches have gained increased attention over the past few years. These models have been used in various applications: image captioning (You et al., 2016); sentiment analysis (Poria et al.,

Proceedings of the 27th International Conference on Computational Linguistics, pages 3793–3805
Santa Fe, New Mexico, USA, August 20-26, 2018.

2016); multimodal machine translation (Specia et al., 2016); visual question answering (Antol et al., 2015); combating human trafficking (Tong et al., 2017); and detection of Cyber-bullying (Zhong et al., 2016). In today's online environment, the strategic use of multimedia (image, video etc.) has become increasingly important part of creating a successful website. Visual information embedded in web documents provides right-angled information that is free of ambiguities of natural language. It can also improve search ranking and Search engine optimization (SEO) scores on many levels that contribute to search visibility, find-ability, user satisfaction, experience and engagement. As a result, use of multimedia content in web documents has become more dominant than text in recent years. Rich content of multimedia data, constructed in alliance with the information contained in different modalities, calls for new and innovative methods for better management of web search results.

In this paper, we hypothesize that high quality clustering can be obtained by representing different objective functions over different views of web documents and simultaneously optimizing them. In contrast to prior works which solely depend on information extracted from text we present a multi objective based multiview clustering algorithm which integrates both visual (images) and text content of web-documents for solving SRC problem. Experimental results show that this new multi-view approach significantly outperforms state-of-the-art unimodal approaches. We motivate this paper with a real example which

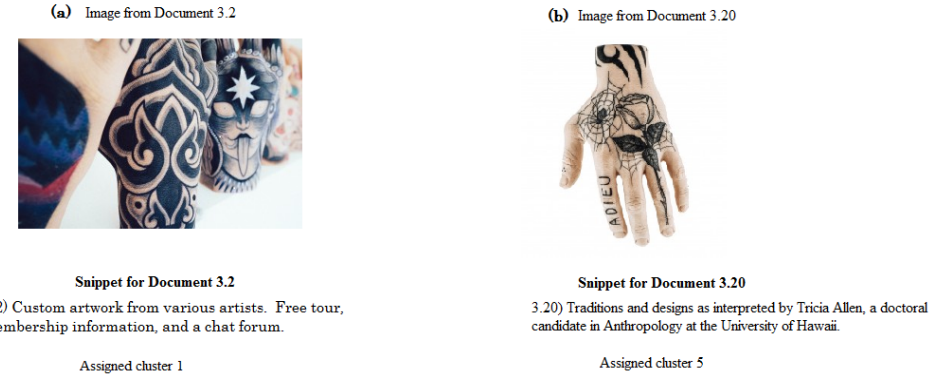


Figure 1: Example where text content differs but image information is same.

demonstrates the potential benefit of integrating visual information for better organization of web search results. Figure 1 shows results from state-of-the-art approach that depends only on text part of the web document for the web search query "3.1 Arts_Bodyart_Tattoo" in ODP-239 dataset. Even though it is clear from looking at the images that they should belong to same cluster, text snippets are not providing enough clues, causing it to appear in different clusters. It is evident from the example that there is complimentary information in the images that is either unavailable in the text or can be in contrast to the text. Our hypothesis is that we can improve SRC performance if we can capture this complimentary high-level visual information. We use pre-trained 19-layer VGG net (Simonyan and Zisserman, 2014) to extract high-level features directly from the image. After extracting the image features for all the images in our dataset, *image view* of each document is generated (Section 3.1). To generate the *textual view* of each document, we combine the benefits of both Word2Vec (Mikolov et al., 2013) and TF-IDF. While Word2Vec vectors are better at capturing the generalized meaning of each word, combining them together by assigning equal weight to all words of a document to generate its *textual view* is not ideal for our task. For example, words that contribute to the syntax rather than the meaning of a sentence should have lower impact on clustering algorithm compared to more specific rare words. Therefore, we scale each word vector by the corresponding TF-IDF weight for that word and generate the *textual view* following the approach in Section 3.1. Finally, we combine both views of the dataset by our multi-objective based multi-view clustering framework.

In order to draw conclusive remarks, we present an exhaustive evaluation where our multi objective based multi-view algorithm (MOO-Multiview-PBM) is compared to the most competitive text-based (endogenous) SRC algorithms: STC (Zamir and Etzioni, 1998), Bisecting Incremental K-means (BIK), LINGO (Osinski and Weiss, 2005), GK-means (Moreno et al., 2013), MOO-clus (Acharya et al., 2014)

and MMOEA (Wahid et al., 2014). Experiments are conducted on three different standard data sets (MORESQUE, ODP- 239 and AMBIENT) for two clustering evaluation metrics (F_{b^3} and F_1). Results show that MOO-Multiview-PBM exceeds all text-based approaches and solutions. In this paper, our main contributions are as follows: **a)** As far the best of our knowledge, this is the first attempt to solve SRC by using both textual & visual information; **b)** A new multi objective based multi-view clustering algorithm for SRC, that determines the number of clusters automatically; and **c)** Novel text representation of document in the context of SRC by combining the benefits of both word2vec and TF-IDF.

2 Related Work

2.1 SRC Algorithms

Suffix Tree Clustering (STC) algorithm proposed by Zamir and Etzioni (1998), is a clustering technique, that combines base clusters having maximum string overlaps based on web snippets represented as compact tries. Results showed improvements over K-means, agglomerative hierarchical clustering, buckshot, single-pass and fractionation algorithms, and this approach is a tough baseline to beat Moreno and Dias (2014). Later, authors of Osinski and Weiss (2005) presented an approach named LINGO which utilizes similar representation of strings as done in Zamir and Etzioni (1998). Initially frequent phrases were extracted based on suffix-arrays and later the group descriptions were matched with topics generated with latent semantic analysis. Documents were then assigned to their relevant groups. Carpineto et al. (2009) showed that the nature of the outputs obtained by SRC algorithms recommend the adoption of a meta clustering approach. The core idea is to combine the complementary results obtained from SOO (single objective) solutions. A novel approach that computes the agreement of two partitions of objects into varied clusters was proposed in this paper. This is done depending on the information content related to the series of decisions made by the partitions on single pairs of objects. OPTIMSRC results demonstrated that meta clustering is far better than individual clustering techniques. Moreno et al. (2013), adapted the K-means algorithm to a third-order similarity measure and proposed a stopping criterion that determines the optimal number of clusters automatically. Experiments were conducted on two standard data sets, MORESQUE (Navigli and Crisafulli, 2010) and ODP-239 (Carpineto and Romano, 2010), and showed significant improvement over all existing text-based SRC techniques developed by then.

Later, Acharya et al. (2014), first defined the SRC task as a multi-objective problem. They defined two objective functions (separability & compactness), that are optimized parallely with the help of AMOSA (Bandyopadhyay et al., 2008). Their evaluations outperformed knowledge-driven exogenous strategies (Scaiella et al., 2012), text-based endogenous SRC approaches and algorithms.

In another work, the multi-view clustering approach proposed by Wahid et al. (2014) used the search capability of multi-objective optimization for SRC. It is basically a cluster ensemble approach where the outputs of multiple clustering techniques like hierarchical clustering approaches and K-means are combined efficiently using NSGA-II (Deb et al., 2002) (non-dominated sorting genetic algorithm-II).

A considerable amount of works have also been proposed that uses exogenous information to solve the SRC problem. One such work is proposed by Scaiella et al. (2012) which uses Wikipedia articles to develop a bipartite graph and employs spectral clustering over it to discover relevant clusters. Recently, authors of Di Marco and Navigli (2013) presented an approach to incorporate word sense induction on the Web1T corpus (Brants and Franz, 2006) that improves SRC.

2.2 Multi-view Clustering

Multi-view data sets are frequent in real life because of the use of different modalities of data input and generation, viz., text, video and audio. The growth of multi-view data in real-world applications has increased the curiosity in multi-view learning (Sun, 2013). Multi-view clustering techniques try to explore the available multiple representations of data for obtaining a precise and robust partitioning of the data in contrast to single-view clustering. A two-view expectation maximization based clustering technique was developed by (Bickel and Scheffer, 2004). A two-view spectral clustering algorithm that generates a bipartite graph was developed by (De Sa, 2005). Another multi-view spectral clustering technique was proposed by (Kumar and Daumé, 2011). A convex mixture model based multi-view

clustering technique was proposed by (Tzortzis and Likas, 2009). The same authors later proposed a kernel-based weighted multi-view clustering technique (Tzortzis and Likas, 2012). A cluster ensemble based approach for multi-view clustering was proposed by (Xie and Sun, 2013). A new multi-view based K-means clustering technique was developed by (Cai et al., 2013) for large-scale data sets. (Wahid et al., 2014) have developed a cluster ensemble based technique to solve the multi-view clustering problem for web documents. Although it presented a multi-view based algorithm MMOEA for solving SRC problem but all the views (i.e., different representations of the data set) are related to semantic and syntactic contents of the web snippets.

3 Multiobjective Multi-view Approach for SRC

In this work we have proposed a multiobjective based multi-view approach, namely *MOO-Multiview-PBM* for solving the problem of search result clustering.

3.1 Generation of Different Views

In the current study we have used some standard data sets of SRC problem for the purpose of evaluation: AMBIENT, MORESQUE (Navigli and Crisafulli, 2010) and ODP-239 (Carpineto and Romano, 2010). For each of web-snippets present in the data set, we have generated two views as follows:

1. *The Textual view*: This view represents both syntactic and semantic information of a document given a particular query. This is generated by combining the document similarity matrix obtained from both word embedding and TF-IDF. Using word embedding each word in the vocabulary is represented by a vector of dimension 1×100. Document similarity matrix using word embedding is generated by Equation 1.

$$S_{emb}(\overline{d}_i, \overline{d}_j) = \frac{1}{\|d_i\|\|d_j\|} \sum_{r=1}^{\|d_i\|} \sum_{b=1}^{\|d_j\|} CosSim(w_i^r, w_j^b) \tag{1}$$

   Here, w_i^r (resp. w_j^b) represents the r^{th} (resp. b^{th}) word vector of document $\overline{d}_i$ (resp. $\overline{d}_j$). $\|d_i\|$ and $\|d_j\|$ denote the total number of words in documents $\overline{d}_i$ and $\overline{d}_j$, respectively. $CosSim(.,.)$ is the cosine similarity between two vectors. TF-IDF is generated using the following steps:
   (a) The terms in the documents are first extracted,
   (b) A document-term matrix is created. Here, a row, a column and a cell correspond to a document, a term, and the weighted value of a term for a document, respectively,
   (c) TF-IDF (a common weighting scheme) values are used to fill this matrix. Each cell contains the TF-IDF score of a term given a document.

   The cosine similarity is calculated between TF-IDF vectors of two documents to generate the $document \times document$ similarity matrix S_{tfidf}. Here $S_{tfidf}(\overline{d}_i, \overline{d}_j)$ contains cosine similarity between TF-IDF vectors of two documents, $\overline{d}_i$ and $\overline{d}_j$.
   The Final $document \times document$ similarity matrix, S_{text}, is generated by the equation:

$$S_{text}(\overline{d}_i, \overline{d}_j) = S_{emb}(\overline{d}_i, \overline{d}_j) \times S_{tfidf}(\overline{d}_i, \overline{d}_j), i, j = 1, \ldots, n \tag{2}$$

   Here n is the total number of documents.
2. *The Image view*: Images are extracted from each web document. We feed images to the pre-trained VGG19 network which computes a 4096 dimensional feature vector for every image that contains the activations of the hidden layer ('fc7') immediately before the VGG's object classifier. Given two image vectors img_i and img_j from two different web documents $\overline{d}_i$ and $\overline{d}_j$, respectively, the similarity between the documents is calculated by Equation 3.

$$S_{image}(\overline{d}_i, \overline{d}_j) = \frac{1}{\|d_i\|\|d_j\|} \sum_{r=1}^{\|d_i\|} \sum_{b=1}^{\|d_j\|} CosSim(img_i^r, img_j^b) \tag{3}$$

   Here, img_i^r (resp. img_j^b) represents the r^{th} (resp. b^{th}) image of the $\overline{d}_i$ (resp. $\overline{d}$) web document. $\|d_i\|$ and $\|d_j\|$ denote the total number relevant images present in $\overline{d}_i$ and $\overline{d}_j$, respectively.

3.2 String Representation and Archive Initialization

The first step of the proposed clustering approach is to initialize the Archive used in AMOSA (Bandyopadhyay et al., 2008) with some alternative diverse set of solutions. The solutions are generated randomly. Here each solution contains a set of cluster centroids (representative web-snippets) in order to represent the partitioning of web-snippets.

The number of cluster centroids encoded in a particular solution i, denoted by K_i, is selected randomly from the given range K_{min} to K_{max} as follows: $K_i = (rand()\text{mod}(K_{max} - 1)) + K_{min}$. For the purpose of initialization, K_i number of web-snippets are randomly selected from the data set and the corresponding indices are used as the initial cluster centers.

3.3 Formation of Clusters and Objective Function Calculations

After initializing the archive members with some randomly selected cluster centroids, the following steps are executed to compute different objective functions. The search capability of AMOSA can be utilized to simultaneously optimize these objective functions.

1. First, the set of representative web-snippets present in the string are extracted. Let the entire set be $\{\overline{C}_1, \overline{C}_2, \ldots, \overline{C}_K\}$ here $\overline{C}_i$ is the ith cluster representative. K is the number of clusters encoded in that particular string. *K*-medoids clustering is applied to the dataset using this set of cluster representatives for different views. In case of text-view, a particular document $\overline{d}_i$ is assigned to cluster t whose centroid has the maximum similarity value to $\overline{d}_i$ given by $t = argmax_{k=1,\ldots K} S_{text}(\overline{d}_i, \overline{C_k})$.Here $S_{text}(\overline{d}_i, \overline{C_k})$ is computed using Equation 2.

   In case of visual view, a particular document $\overline{d}_i$ is assigned to cluster t whose centroid has the maximum similarity value to $\overline{d}_i$. $t = argmax_{k=1,\ldots K} S_{image}(\overline{d}_i, \overline{C_k})$.Here $S_{image}(\overline{d}_i, \overline{C_k})$ is computed using Equation 3.

2. The PBM-index values are calculated for the final partitionings obtained using individual views. Let the values be denoted by PBM_v, $v = 1, 2$.

3. The adjoint matrix (A^v of size $n \times n$, where n is the number of documents) corresponding to view v is calculated as follows:

$$A^v_{ij} = 1 \text{ if } \overline{d}_i \text{ and } \overline{d}_j \text{ belong to the same cluster or } i = \text{j} \quad (4)$$

$$= 0 \quad \text{otherwise} \quad (5)$$

4. A new objective function *Agreement Index* is calculated as follows. This measures the agreement between the partitionings obtained using multiple views. The measure is calculated as follows:

   At a time two views are considered: $v1$ and $v2$. Let the corresponding adjoint matrices be A^{v1} and A^{v2}, respectively. The number of agreement (n_a) is calculated as follows: $n_a = \sum_{i=1}^{n} \sum_{j=1}^{n} I_{A^{v1}_{ij}, A^{v2}_{ij}}$, here

$$I_{A^{v1}_{ij}, A^{v2}_{ij}} = 1 \text{ if } A^{v1}_{ij} = A^{v2}_{ij} \quad (6)$$

$$= 0 \quad \text{otherwise}$$

   The number of disagreements (n_d) is calculated as follows: $n_d = n^2 - n_a$. *Agreement index* between these two views ($v1$, $v2$) is calculated as follows: $AI_{v1,v2} = \frac{n_a+1}{n_d+1}$. The values of 1 in the numerator and denominator are used as a normalization factor to avoid the problem of division by zero. The total *Agreement index* for the entire partitioning is calculated as follows:

$$AI = \frac{\sum_{j=1}^{m} \sum_{i=1, j \neq i}^{m} 2 \times AI_{v_j, v_i}}{m \times (m-1)}, \quad (7)$$

   where m is the available number of views.

5. The objective functions corresponding to a particular string are:
   $\{PBM_{text}, PBM_{image}, AI\}$ where PBM_{text} and PBM_{image} are the values of PBM-indices calculated on partitionings obtained using text based view and image based view, respectively.

3.4 Update of String

After the objective functions are calculated, a consensus partitioning is obtained which satisfies all the available views. The cluster centroids corresponding to this consensus partitioning are used to update the string of AMOSA.

- Let the partitioning obtained using two views be represented by π^1, π^2. Let us denote the jth cluster of view v as π_j^v. First some reordering is done among all the obtained partitionings so that there is a one-to-one correspondence between the cluster numbers of different partitionings.
- New cluster centroids $\overline{C}$, are selected among those documents which are present in the same cluster corresponding to both the views. Let common documents of cluster i in both the views be $\{\overline{d}_1^i, \ldots, \overline{d}_m^i\}$. Then the representative of this set, denoted by $\overline{C}_i$, is $\overline{d}_t$ where
$$t = argmax_{k=1}^{m} \frac{\sum_{j=1, j\neq k}^{m} \frac{S_{text}(\overline{d}_k^i, \overline{d}_j^i) + S_{image}(\overline{d}_k^i, \overline{d}_j^i)}{2}}{m-1}. \quad (8)$$
Here S_{final} and S_{image} are calculated using Equation 2 and 3, respectively.
- Next, the newly generated cluster centroids $\overline{C}_j, \quad j = 1, \ldots, K$ are used to obtain the final consensus partitioning as follows: $\pi_j = \{\forall \overline{d}_i \in S : Sim(\overline{d}_i, \overline{C}_j) > Sim(\overline{d}_i, \overline{C}_l) \text{ for } l = 1, \ldots, K, l \neq j\}$. Here, K is the number of clusters encoded in that solution.
$$Sim(\overline{d}_i, \overline{C}_j) = \frac{S_{text}(\overline{d}_i, \overline{C}_j) + S_{image}(\overline{d}_i, \overline{C}_j)}{2} \quad (9)$$
Here S_{final} and S_{image} are calculated using Equation 2 and 3, respectively. S denotes the set of all documents.
- Finally, the representatives of clusters $\pi_j, \quad j = 1 \ldots, K$ are calculated using Equation 8. These new cluster centroids $\overline{C}_j, \ j = 1, \ldots, K$ are used to replace the old centroids encoded in the string.

So, in order to obtain a consensus partitioning, initially the common points of different clusters present in different partitionings obtained using different views are identified. These points are further used to determine cluster centroids. The other points are assigned to the centroids having maximum similarity, calculated by Equation 9, to obtain a final consensus partitioning. These cluster centroids are used to update the given string.

3.5 Search Operators

In order to explore the search space efficiently using AMOSA, perturbation operations are introduced. These operators also help in generating some new solutions from the current solution which can further take part in the search process. For this purpose, three different perturbation operators are introduced. Below we describe three types of mutation operations in detail:

Mutation 1: In this operation each centroid is parsed individually and with some probability the existing document id in the centroid is replaced by a new document id which is selected from the document collection randomly [1].

Mutation 2: Here a cluster centroid is randomly selected and it is deleted from the string. Size of the string is reduced by 1.

Mutation 3: Here a web document is randomly chosen and the corresponding index is added to the string. Size of the string is increased by 1.

All the above mentioned three mutation operations are equi-probable. Any one of the above discussed mutation operators is applied on a particular solution to generate a new solution which can further participate in the process of AMOSA.

4 Results and Discussion

Results of different approaches are shown in Table 2 for all the data sets with respect to different performance metrics. From Table 2, it is evident that our proposed approach *MOO-Multiview-PBM* outperforms the results of *MOO-clus* (Acharya et al., 2014) by a margin of around $6.2\% - 6.5\%$ in terms

[1] existing document id may or may not get replaced by the new document id

of $F_{b^3} - measure$ and $7.0\% - 7.3\%$ in terms of $F1 - measure$ for both MORESQUE and ODP-239 datasets. Box plots of results obtained from combined $F1 - measure$ and combined $RandIndex$ over all three data sets (i.e., AMBIENT, MORESQUE and ODP-239) by our proposed method are reported in Figure 4. This is done to compare the performance of *MOO-Multiview-PBM* with MMOEA algorithm (Wahid et al., 2014). Although, AMBIENT has received less attention since the creation of ODP-239, we have included it to show a fair comparison of our results with the results from MMOEA (Wahid et al., 2014). In Figure 4, comparison with MMOEA (Wahid et al., 2014) is illustrated. *MOO-Multiview-PBM (word2vec*tfidf)* reports an average F1 measure and Rand Index values of 0.74 and 0.77, respectively, over all three datasets combined, which are similar to those obtained by MMOEA. Note that MMOEA is based on three views (Topics, terms and senses) and also utilized some external information like Wikipedia data during its processing. But our proposed method attains comparable results using information extracted only from the given datasets. No external information was used during the computation of our approach.

Table 1: Evaluation results in terms of $F1$ and F_{b^3} over MORESQUE and ODP239 data sets: Comparison of the proposed approach with state-of-the-art approaches.†→ Results are obtained by 10 consecutive runs of the algorithm and are statistically significant.

		MOO-Multiview-PBM (word2vec*tf-idf)		**MOO-Multiview-PBM (word2vec)**		**MOO-Clus**		**SOO-SRC**			
		Min	*Max*	*Min*	*Max*	*Min*	*Max*	*GK-means*	*STC*	*LINGO*	*BIK*
MORESQUE	F_{b^3}	0.506	**0.564**†	0.510	0.557	0.477	0.502	0.482	0.460	0.399	0.315
	F1	0.698	**0.742**†	0.682	0.728	0.658	0.675	0.655	0.455	0.326	0.317
ODP-239	F_{b^3}	0.491	**0.549**†	0.482	0.531	0.478	0.484	0.452	0.403	0.346	0.307
	F1	0.438	**0.474**†	0.431	0.462	0.379	0.384	0.366	0.324	0.273	0.2

Table 2: Evaluation results in terms of $F1$ and F_{b^3} over MORESQUE and ODP239 data sets: Comparison of the proposed approach over each single view.

		MOO-image		**MOO-word2vec**		**MOO-word2vec*tfidf**	
		Min	*Max*	*Min*	*Max*	*Min*	*Max*
MORESQUE	F_{b^3}	0.427	0.4684	0.479	0.5174	0.489	0.5267
	F1	0.613	0.657	0.664	0.6812	0.6793	0.6973
ODP-239	F_{b^3}	0.4429	0.4725	0.4803	0.4831	0.4821	0.4921
	F1	0.342	0.376	0.3814	0.3901	0.4031	0.4083

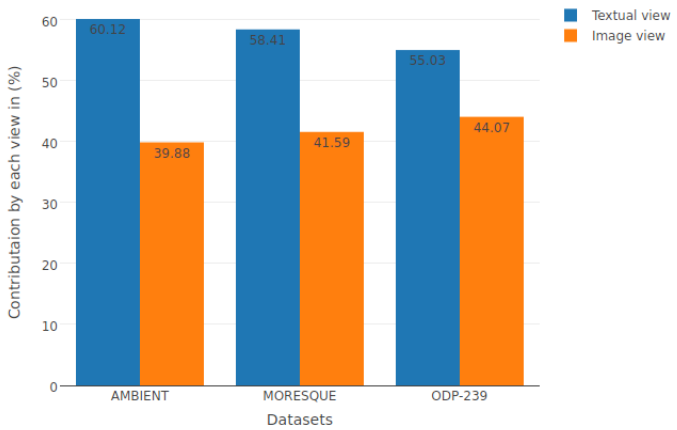


Figure 2: The following bar graph shows contribution by each view in the clustering process for different datasets.

We have implemented our proposed algorithm using two different versions of textual views. Initially we have used word embeddings (Mikolov et al., 2013) for the textual view and combined it with image view in our algorithm *MOO-Multiview-PBM(word2vec)*. Later we have used word embeddings and tf-idf vector together, described in Section 3.1, as textual view along with image view for our algorithm *MOO-Multiview-PBM(word2vec*tf-idf)*. Results in Table 2 show that *MOO-Multiview-PBM(word2vec*tf-idf)* achieves an improvement of $1.2\% - 1.8\%$ over *MOO-Multiview-PBM(word2vec)*. It substantiates our hypothesis that combining the benefits of both word2vec and TF-IDF can improve the search result clustering performance.

We have also quantified the influence of individual views in obtaining the final partitioning. The degree is expressed as follows:

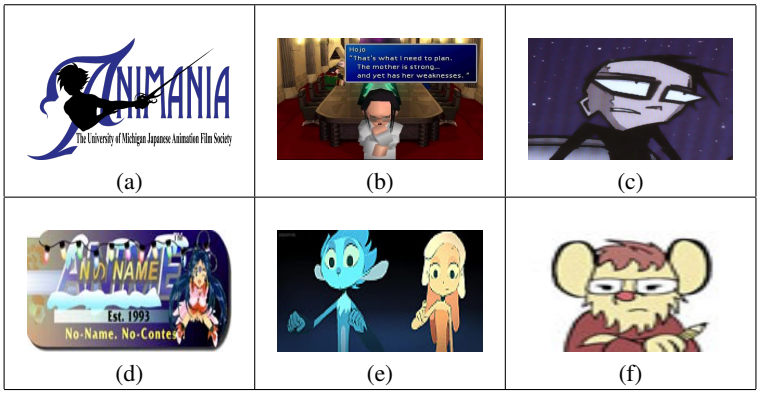
$$Degree_v = \frac{\|A^v \cap A^{\cup v}\|}{\|A^{\cup v}\|}$$

where A^v denotes the adjoint matrix (expressed in Equation 4) corresponding to the partitioning obtained using view v and $A^{\cup v}$ is the adjoint matrix corresponding to the consensus partitioning obtained by consulting all views. The contributions computed as above for different views are shown in Figure 2. It is observed that knowledge extracted from images has contributed on an average 41.58% in the clustering process (refer to Figure 2) for all three datasets.

As mentioned in Section 1 that only textual information is not sufficient for SRC clustering, here we have presented some examples where although these web snippets belong to the same cluster, their textual information widely vary whereas their corresponding image information classify them into the same cluster. In Table 3, we have listed some sample queries from ODP-239 datasets whose textual contents vary widely related to query. In Figure 3, we have shown corresponding images extracted from websites of above mentioned queries. It is evident from these examples that image information is more relevant to the query compared to the textual information.

Table 3: List of query results whose textual contents differ. **S#Id**: Subtopic Id. **R#Id**: Result Id.

S#Id	Subtopic	R#Id	Results	Figures
1.1	Arts	1.19	List of screenings, a few MP3s, and series information.	3a and 3d
	Animation	1.20	Events, news, forum, and newsletter.	3b and 3e
	Anime	1.21	History, constitution, showings, and tape library.	3c and 3f

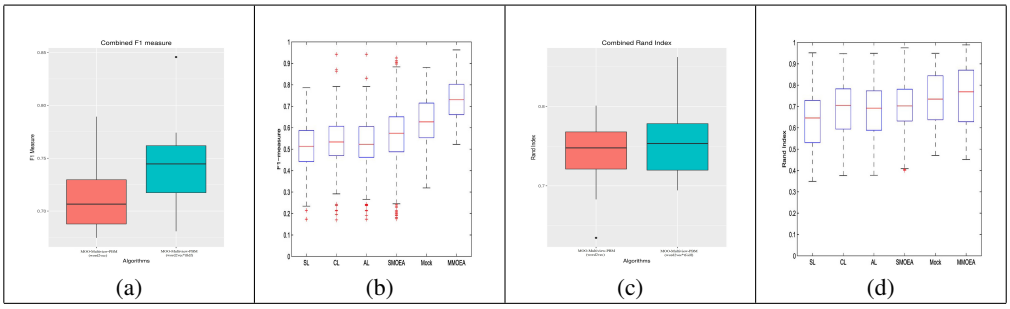


(a) (b) (c) (d) (e) (f)

Figure 3: Sample figures extracted from each query given in Table 3.

5 Error Analysis

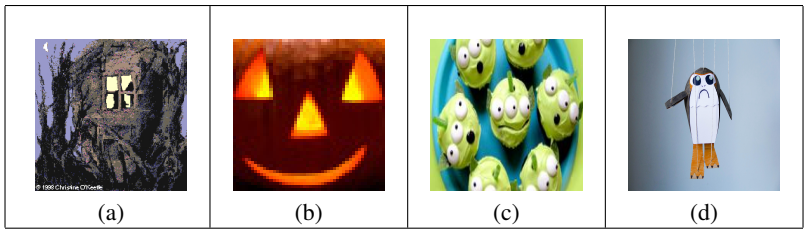
The errors made by our proposed method have been thoroughly analyzed. After a thorough manual analysis, it has been observed that misclassification occurred because of the following reasons. Firstly, from Table 4, it can be seen that there are many inactive web links from which we failed to extract images. For those instances only textual information from the snippets are used for classification. For example, in ODP-239 dataset results 2.54 and 2.65 belong to the same cluster in original labelling. Web links of both 2.54 and 2.65 are inactive, therefore only textual information is used for clustering.



(a) (b) (c) (d)

Figure 4: Boxplots of the best (a) F1-measure (c) Rand index values obtained for different queries combining all the data sets together after application of the proposed MOO-Multiview-PBM technique. Boxplots of the best (b) F1-measure (d) Rand index values obtained for different queries combining all the data sets together after application of MMOEA (taken from the paper (Wahid et al., 2014)).

But in these web snippets text varies widely, viz.,"Meeting schedules, contact information, calendar of events, ftm, mtf, soffa, information, articles, book catalog, and guestbook." in 2.65 and "Headline links from media sources worldwide." in 2.54, hence in predicted clustering solution both 2.54 and 2.65 belong to different clusters. Secondly, there are few instances where textual view and image view differ widely, consensus between these two views is very less hence leads to misclassification. For example, in ODP-239 dataset query 196.1 and 196.11 originally belong to the same cluster. But the text contents, viz.,"Fairy and monster identification list, free e-cards, customs, recipes, games, history, and links to other holiday pages." in 196.1 and "Includes crafts, recipes, costumes, games and activities, and party ideas. Also features autumn harvest party ideas with pumpkin crafts." in 196.11, and image contents of queries 196.1 (see Figures 5a and 5b) and 196.11 (see Figures 5c and 5d) differ widely hence no concrete consensus is drawn between the two views, therefore in predicted clustering solution they are misclassified.



(a) (b) (c) (d)

Figure 5: Sample figures extracted from queries in dataset ODP-239.

6 Conclusion

In the current study a multiobjective based multi-view clustering technique is developed for solving the problem of search result clustering (SRC). Views constructed over textual and visual information contained in web-snippets are considered. We have hypothesized that two web-snippets can be similar either with respect to content or with respect to images. These two views are exploited simultaneously in order to detect good-quality clustering of web-snippets. Three objective functions capturing the qualities of different partitions and a consensus function measuring the similarity between two partitions obtained using different views are simultaneously optimized using the search capability of a MOO process. Improved results on standard bench-mark data sets over state-of-the-art approaches support our hypothesis that use of multi-view information indeed helps in solving the SRC problem. In future we would like to exploit other views of web-snippets and incorporate it in our framework. Investigations of AMOSA as the underlying optimization strategy and PBM -index as the internal cluster validity index are also required to be carried out in future.

Acknowledgements

Mohammed Hasanuzzaman and Andy Way would like to acknowledge ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-

funded under the European Regional Development Fund.

References

Sudipta Acharya, Sriparna Saha, Jose G Moreno, and Gaël Dias. 2014. Multi-objective search results clustering. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 99–108.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. A general evaluation measure for document organization tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 643–652. ACM.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Sanghamitra Bandyopadhyay, Sriparna Saha, Ujjwal Maulik, and Kalyanmoy Deb. 2008. A simulated annealing-based multiobjective optimization algorithm: Amosa. *IEEE transactions on evolutionary computation*, 12(3):269–283.

Steffen Bickel and Tobias Scheffer. 2004. Multi-view clustering. In *ICDM*, volume 4, pages 19–26.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1.

Xiao Cai, Feiping Nie, and Heng Huang. 2013. Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604.

Claudio Carpineto and Giovanni Romano. 2010. Optimal meta search results clustering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177. ACM.

Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. 2009. A survey of web clustering engines. *ACM Computing Surveys (CSUR)*, 41(3):17.

Daniel Crabtree, Xiaoying Gao, and Peter Andreae. 2005. Improving web clustering by cluster selection. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 172–178. IEEE Computer Society.

Virginia R De Sa. 2005. Spectral clustering with two views. In *ICML workshop on learning with multiple views*, pages 20–27.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197.

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

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Abhishek Kumar and Hal Daumé. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 393–400.

Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Anirban Mukhopadhyay. 2011. *Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics*. Springer Science & Business Media.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jose G Moreno and Gaël Dias. 2014. Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In *14th Conference of the European Chapter of the Association for Computational Linguistics*.

José G Moreno, Gaël Dias, and Guillaume Cleuziou. 2013. Post-retrieval clustering using third-order similarity measures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 153–158.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 116–126. Association for Computational Linguistics.

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

Malay K Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3):487–501.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

Ugo Scaiella, Paolo Ferragina, Andrea Marino, and Massimiliano Ciaramita. 2012. Topical clustering of search results. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 223–232. ACM.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 543–553.

Shiliang Sun. 2013. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1547–1556.

Grigorios Tzortzis and Aristidis Likas. 2009. Convex mixture models for multi-view clustering. In *International Conference on Artificial Neural Networks*, pages 205–214. Springer.

Grigorios Tzortzis and Aristidis Likas. 2012. Kernel-based weighted multi-view clustering. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 675–684. IEEE.

Abdul Wahid, Xiaoying Gao, and Peter Andreae. 2014. Multi-view clustering of web documents using multi-objective genetic algorithm. In *IEEE Congress on Evolutionary Computation*, pages 2625–2632.

Xijiong Xie and Shiliang Sun. 2013. Multi-view clustering ensembles. In *Machine Learning and Cybernetics (ICMLC), 2013 International Conference on*, volume 1, pages 51–56. IEEE.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Oren Zamir and Oren Etzioni. 1998. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54. ACM.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958.

Appendix A. Objective Functions

PBM index: This is a popular cluster validity index proposed by Pakhira, Bandyopadhyay and Maulik (Pakhira et al., 2004). It outperforms most of the cluster validity indices in the literature in properly detecting the optimal partitioning. This index is defined as follows:

$$PBM(K) = (\frac{1}{K} \times \frac{\mathcal{E}_1}{\mathcal{E}_\mathcal{K}} \times D_K) \tag{10}$$

Here, K denotes the number of clusters, $\mathcal{E}_\mathcal{K} = \sum_{k=1}^{K}\sum_{j=1}^{n_k} dist(\overline{C}_k, \overline{d}_j^k)$ and $D_K = \max_{i,j=1}^{K} dist(\overline{C}_i, \overline{C}_j)$, where $\overline{C}_j$ denotes the centroid of the j^{th} cluster and $\overline{d}_j^k$ denotes the j^{th} web-snippet of the k^{th} cluster. n_k is the total number of web-snippets of the k^{th} cluster. The objective is to maximize the PBM-index. In case of text-view, $dist(\overline{C}_k, \overline{d}_j^k) = 1 - S_{text}(\overline{C}_k, \overline{d}_j^k)$, where S_{text} is calculated using Equation 2 and in case of image-view $dist(\overline{C}_k, \overline{d}_j^k) = 1 - S_{image}(\overline{C}_k, \overline{d}_j^k)$, where S_{image} is calculated using Equation 3. Similarly D_K value between two centroids is also calculated either using text-based similarity or image based similarity measure.

Appendix B. Experimental Setup

We perform web snippet tokenization and word vector generation using gensim library.[2] Word embeddings are obtained from the pre-trained Google news word embeddings.[3] Image features are extracted from the *FC7* layer of *VGG19* available in Keras library.[4] We executed our algorithm over three gold standard datasets: AMBIENT[5]; MORESQUE (Navigli and Crisafulli, 2010); and ODP-239 (Carpineto and Romano, 2010). Description of the datasets, total number of relevant images extracted for each query and number of active query links present in each data set are summarized in Table 4.

The parameters of our proposed algorithms are: $T_{min} = 0.01$, $T_{max} = 100$, $\alpha = 0.85$, $HL = 20$, $SL = 30$, $itr = 20$, $K_{max} = \sqrt{\#ofsamples}$ and $K_{min} = 2$. All these values have been determined after conducting a thorough sensitivity study (and those are in line with the approach proposed by (Acharya et al., 2014)).

Table 4: First part represents SRC gold standard data sets. Second part represents total number of relevant images extracted for each query and number of active query links present in each data set.

Datasets	**# of quaries**	**# of Subtopics Avg/Min/Max**	**# of Snippets**
AMBIENT	44	17.95/6/37	4400
MORESQUE	114	10 / 10 / 10	11400
ODP-239	239	6.7 / 2 / 38	25580

# of web links	**# of active web links**	**# of inactive web links**	**# of images in each quary Avg/Min/Max**
4400	4137	263	6 / 4 / 20
11400	10834	566	8/ 4 / 18
25580	19513	3067	6 / 3 / 25

Appendix C. Evaluation Metrics

An ideal SRC system should be represented by a unique cluster having all the relevant web pages inside. However, determining a unique and complete metric to evaluate the performance of a clustering algorithm is still an open problem (Amigó et al., 2013).

In order to measure the qualities of partitions obtained using different clustering techniques for these web search data sets, we have used three cluster quality measures, F_{b^3} measure (Amigó et al., 2009), F_1 measure (Crabtree et al., 2005) and Rand Index (Hubert and Arabie, 1985). In particular, F_{b^3} has been defined to evaluate completeness, cluster homogeneity, size-vs-quantity and rag-bag constraints. F_{b^3} is a function of $Precision_{b^3}(P_{b^3})$ and $Recall_{b^3}$ (R_{b^3}). All metrics are defined as follows:.

$$F_{b^3} = \frac{2 \times P_{b^3} \times R_{b^3}}{P_{b^3} + R_{b^3}}, P_{b^3} = \frac{1}{N}\sum_{j=1}^{K}\sum_{d_i \in \pi_j} \frac{1}{\mid \pi_i \mid} \sum_{d_l \in \pi_j} h^*(d_j, d_l),$$

$$R_{b^3} = \frac{1}{N}\sum_{j=1}^{K}\sum_{d_i \in \pi_j^*} \frac{1}{\mid \pi_i^* \mid} \sum_{d_l \in \pi_j^*} h^*(d_j, d_l)$$

here π_j is th j^{th} cluster and π_j^* is the gold standard of j^{th} cluster. $h(.,.)$ and $h^*(.,.)$ are defined in Equation 11.

[2] https://radimrehurek.com/gensim/
[3] https://code.google.com/archive/p/word2vec/
[4] https://keras.io/
[5] http://credo.fub.it/ambient

$$h^*(d_j, d_l) = \begin{cases} 1 & \iff \exists i \quad : d_j \in \pi_i^* \wedge d_l \in \pi_i^* \\ 0 & \quad\quad\;\; : otherwise \end{cases}, h(d_j, d_l) = \begin{cases} 1 & \iff \exists i \quad : d_j \in \pi_i \wedge d_l \in \pi_i \\ 0 & \quad\quad\;\; : otherwise \end{cases} \tag{11}$$

Appendix D. Model Comparisons

In order to comprehensively evaluate the performance of our proposed approach *(MOO-Multiview-PBM)*, we listed some strong baseline approaches for comparison.

- **MOO-clus(Acharya et al., 2014):** This algorithm uses archived multi-objective simulated annealing framework to simultaneously optimize two objectives, compactness and separation, for clustering web snippets.
- **GK-means(Moreno et al., 2013):** This algorithm has adapted the K-means algorithm to a third-order similarity measure and proposed a stopping criterion to automatically determine the number of clusters.
- **Suffix Tree Clustering(STC)(Zamir and Etzioni, 1998):** It is an incremental, linear time algorithm which creates clusters based on phrases shared between documents.
- **LINGO(Osinski and Weiss, 2005):** In this method, initially the frequent phrases based on suffix-arrays are extracted and later matched the group description with topics obtained with latent semantic analysis.
- **MMOEA(Wahid et al., 2014):** This algorithm uses multiple views to generate different clustering solutions and then select a combination of clusters to form a final clustering solution.