Multi-Objective Search Results Clustering

Sudipta Acharya **Sriparna Saha** **Jose G. Moreno** **Gaël Dias**
Indian Institute of Technology Patna Normandie University - CNRS GREYC
Kurji, Patna, Bihar, India Caen, France
{sudipta.pcs13,sriparna}@iitp.ac.in first.last@unicaen.fr

Abstract

Most web search results clustering (SRC) strategies have predominantly studied the definition of adapted representation spaces to the detriment of new clustering techniques to improve performance. In this paper, we define SRC as a multi-objective optimization (MOO) problem to take advantage of most recent works in clustering. In particular, we define two objective functions (compactness and separability), which are simultaneously optimized using a MOO-based simulated annealing technique called AMOSA. The proposed algorithm is able to automatically detect the number of clusters for any query and outperforms all state-of-the-art text-based solutions in terms of F_β-measure and F_{b^3}-measure over two gold standard data sets.

1 Introduction

Web search results clustering (SRC), also known as post-retrieval clustering or ephemeral clustering has received much attention for the past twenty years for easing up user's effort in web browsing. The key idea behind SRC systems is to return some meaningful labeled clusters from a set of web documents (or web snippets) retrieved from a search engine for a given query.

Recently, SRC strategies have been focusing on the introduction of external (exogenous) knowledge to better capture semantics between documents (Scaiella et al., 2012; Marco and Navigli, 2013). Although this research direction has evidenced competitive results, the proposed clustering techniques are based on a single cluster quality measure, which must reflect alone the goodness of a given partitioning. These techniques are usually referred to as single objective optimizations (SOO).

In this paper, we hypothesize that improved clustering can be achieved by defining different objective functions over well-known data representations. As such, our study aims to focus on new clustering issues for SRC instead of defining new representation spaces.

Recent studies (Maulik et al., 2011) have shown that clustering can be defined as a multi-objective optimization (MOO) problem. Within the context of SRC, we propose to define two objective functions (compactness and separability), which are simultaneously optimized using a MOO-based simulated annealing technique called AMOSA (Bandyopadhyay et al., 2008).

In order to draw conclusive remarks, we present an exhaustive evaluation where our MOO algorithm (*MOO-clus*) is compared to the most competitive text-based (endogenous) SRC algorithms: STC (Zamir and Etzioni, 1998), LINGO (Osinski and Weiss, 2005), OPTIMSRC (Carpineto and Romano, 2010) and GK-means (Moreno et al., 2013). Experiments are run over two different gold standard data sets (ODP-239 and MORESQUE) for two clustering evaluation metrics (F_β-measure and F_{b^3}-measure). Results show that *MOO-clus* outperforms all text-based solutions and approaches performances of knowledge driven strategies (Scaiella et al., 2012). In this paper, our main contributions are:

- The first[1] attempt to solve SRC by defining multiple objective functions,
- A new MOO clustering algorithm for SRC, which automatically determines the number of clusters,

[1] As far as we know.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 99–108, Dublin, Ireland, August 23-29 2014.

- An exhaustive evaluation of SRC algorithms with recent data sets and evaluation metrics over the most competitive state-of-the-art text-based SRC algorithms.

2 Related Work

2.1 SRC Algorithms

One of the most cited SRC solutions is the Suffix Tree Clustering (STC) algorithm proposed by (Zamir and Etzioni, 1998). They propose a monothetic clustering technique, which merges base clusters with high string overlap based on web snippets represented as compact tries. Their evaluation shows improvements over agglomerative hierarchical clustering, K-Means, Buckshot, Fractionation and Single-Pass algorithms, and is still a hard baseline to beat (Moreno and Dias, 2014).

Later, (Osinski and Weiss, 2005) proposed a polythetic solution called LINGO based on the same string representation as of (Zamir and Etzioni, 1998). They first extract frequent phrases based on suffix-arrays and match group descriptions with topics obtained with latent semantic analysis. Documents are then assigned straightforwardly to their corresponding groups. Their evaluation does not allow conclusive remarks but they propose an open source implementation, which is an important contribution.

More recently, (Carpineto and Romano, 2010) showed that the characteristics of the outputs returned by SRC algorithms suggest the adoption of a meta clustering approach. The underlying idea is that different SOO solutions lead to complementary results that must be combined. So, 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. The results of OPTIMSRC demonstrate that meta clustering is superior over individual clustering techniques.

The latest work, exclusively based on endogenous information (i.e. web snippets returned by the search engine), is proposed by (Moreno et al., 2013). They adapt the K-means algorithm to a third-order similarity measure and propose a stopping criterion to automatically determine the "optimal" number of clusters. Experiments are run over two gold standard data sets, ODP-239 (Carpineto and Romano, 2010) and MORESQUE (Navigli and Crisafulli, 2010), and show improved results over all state-of-the-art text-based SRC techniques so far.

A great deal of works have also proposed to include exogenous information to solve the SRC problem. One important work is proposed by (Scaiella et al., 2012) who use Wikipedia articles to build a bipartite graph and apply spectral clustering over it to discover relevant clusters. More recently, (Marco and Navigli, 2013) proposed to include word sense induction based on the Web1T corpus (Brants and Franz, 2006) to improve SRC. In this paper, we exclusively focus on endogenous solutions.

2.2 MOO-based Clustering

Many works have been proposed where the problem of clustering is posed as one of multi-objective optimization (Deb, 2009; Maulik et al., 2011). One important work is proposed by (Handl and Knowles, 2007) who define a multi-objective clustering technique with automatic K-determination called MOCK. Their algorithm outperforms several standard single-objective clustering algorithms (K-means, agglomerative hierarchical clustering and ensemble clustering) on artificial data sets.

In parallel, a multi-objective evolutionary algorithm for fuzzy clustering is proposed by (Bandyopadhyay et al., 2007) for clustering gene expressions. Here, two objectives are simultaneously optimized. The first one is the objective function optimized in the fuzzy C-means algorithm (Bezdek, 1981) and the other one is the Xie-Beni index (Xie and Beni, 1991).

Later, (Mukhopadhyay and Maulik, 2009) proposed a novel approach that combines the multi-objective fuzzy clustering method of (Bandyopadhyay et al., 2007) with a Support Vector Machines (SVM) classifier. Performance results are provided for remote sensing data.

As far as we know, within text applications, (Morik et al., 2012) is the first work, which formulates text clustering a multi-objective optimization problem. In particular, they express desired properties of frequent termset clustering in terms of multiple conflicting objective functions. The optimization is solved by a genetic algorithm and the result is a set of Pareto-optimal solutions. Note that this effort is

defined for large text colllections with high dimensional data, which is contradictory to the specific task of SRC (Carpineto et al., 2009)[2].

2.3 Our Motivation

Recent works have focused on the introduction of external (exogenous) knowledge to solve the SRC task. However, this research direction higly depends on existing resources, which are not available for a great deal of languages. Moreover, (Carpineto and Romano, 2010) has suggested an interesting research direction, which has still remained unexplored. Indeed, (Carpineto and Romano, 2010) showed that meta clustering leads to improved results in the context of text-based (endogenous) SRC. This suggests that better clustering can be obtained by combining different SOO solutions. However, their algorithm is casted to a SOO problem of the concordance between the clustering combination and a meta partition.

As a consequence, we hypothesize that improved performances can be obtained by defining the SRC task as a MOO clustering problem. For that purpose, we (1) take advantage of the recent advances in the field of multi-objective clustering (Saha and Bandyopadhyay, 2010), (2) define new objective functions in a non euclidean space and (3) adapt a MOO-based simulated annealing technique called AMOSA (Bandyopadhyay et al., 2008) to take into account third-order similarity metrics (Moreno et al., 2013).

3 Clustering as a MOO Problem

3.1 Formal Definition of MOO Clustering

Multi-objective optimization can be formally stated as finding the vector $\overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T$ of decision variables that simultaneously optimize *M* objective function values $\{f_1(\overline{x}), f_2(\overline{x}), \ldots, f_M(\overline{x})\}$ while satisfying user-defined constraints, if any.

An important concept in MOO is that of domination. Within the context of a maximization problem, a solution $\overline{x_i}$ is said to dominate $\overline{x_j}$ if $\forall k \in 1, 2, \ldots, M,\ f_k(\overline{x_i}) \geq f_k(\overline{x_j})$ and $\exists k \in 1, 2, \ldots, M,$ such that $f_k(\overline{x_i}) > f_k(\overline{x_j})$.

Among a set of solutions R, the non-dominated set of solutions $R^{'}$ are those that are not dominated by any member of the set R and is called the globally Pareto-optimal set or Pareto front. In general, a MOO algorithm outputs a set of solutions not dominated by any solution encountered by it. These notions can be illustrated by considering an optimization problem with two objective functions (f_1 and f_2) with six different solutions, as shown in Figure 1. Here target is to maximize both objective functions f_1 and f_2.

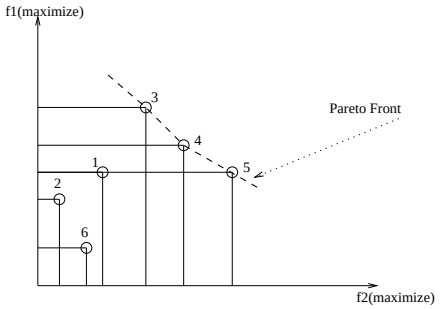

Figure 1: Example of dominance and Pareto optimal front.

In this example, solutions 3, 4 and 5 dominate all the other three solutions 1, 2 and 6. Solutions 3, 4 and 5 are nondominating to each other. Because 3 is better than 4 w.r.t. function f_1, but 4 is better than 3 w.r.t. f_2. Similarly 4 is better than 5 w.r.t. f_1 but 5 is better than 4 w.r.t. f_2. The same happens for solutions 3 and 5. So, the Pareto front is made of solutions 3, 4 and 5.

Within the specific context of clustering, two objective functions are usually defined, which must be optimized simultaneously. These functions are based on two intrinsic properties of the data space and are defined as follows.

[2]SRC is usually referred to as text clustering in the "small": i.e. small list of short text documents.

Compactness: This objective function measures the proximity among the various elements of a given cluster and must be maximized.

Separability: This objective function measures the similarity between two cluster centroids and must be minimized.

3.2 AMOSA Optimization Strategy

Clustering is viewed as a search problem, where optimal partitions satisfying the given set of objective functions must be discovered. As such, an optimization strategy must be defined. Here, we propose to use archived multi-objective simulated annealing (AMOSA) proposed by (Bandyopadhyay et al., 2008). AMOSA incorporates the concept of an archive where the non-dominated solutions seen so far are stored.

Two limits are kept on the size of the archive: a hard limit denoted by *HL* and a soft limit denoted by *SL*. Given $\gamma > 1$, the algorithm begins with the initialization of a number $(\gamma \times SL)$ of solutions each of which representing a state in the search space. Thereafter, the non-dominated solutions are determined and stored in the archive.

Then, one point is randomly selected from the archive. This is taken as the current point, or the initial solution, at temperature $T = T_{max}$. The current point is perturbed/mutated to generate a new solution named new-pt and its objective functions are computed. The domination status of the new-pt is checked w.r.t. the current point and the solutions in the archive. Based on domination status, different cases may arise: (i) accept the new-pt, (ii) accept the current-pt or (iii) accept a solution from the archive. In case of overflow of the archive, its size is reduced to *HL*.

The process is repeated *iter* times for each temperature that is annealed with a cooling rate of α ($<$1) till the minimum temperature T_{min} is attained. The process thereafter stops and the archive contains the final non-dominated solutions i.e. the Pareto front.

4 SRC as MOO Problem: MOO-clus

4.1 Archive Initialization

As we follow an endogenous approach, only the information returned by a search engine is used. In particular, we only deal with web snippets and each one is represented as a word feature vector. So, our solution called *MOO-clus* starts its execution after initializing the archive with some random solutions as archive members. Here, a particular solution refers to a complete assignment of web snippets (or data points) in several clusters. So, the first step is to represent a solution compatible with AMOSA, which represents each individual solution as a string. In order to encode the clustering problem in the form of a string, a center-based representation is used. Note that the use of a string representation facilitates the definition of individuals and mutation functions (Bandyopadhyay et al., 2008).

Let us assume that the archive member i represents the centroids of K_i clusters and the number of tokens in a centroid is p[3], then the archive member (or string) has length l_i where $l_i = p \times K_i$. To initialize the number of centroids K_i encoded in the string i, a random value between 2 and K_{max} is chosen and each K_i cluster centroid is initialized by randomly generated tokens from the global vocabulary.

4.2 Assignment of Web Snippets

As for any classical clustering algorithms, web snippets (or data points) must be assigned to their respective clusters. In *MOO-clus*, this assignment is computed as in (Moreno et al., 2013), to take advantage of recent advances in similarity measures. For two word feature vectors d_i and d_j, their similarity is evaluated by the similarity of their constituents as defined in Equation 1.

$$S(d_i, d_j) = \frac{1}{\|d_i\|\|d_j\|} \sum_{r=1}^{\|d_i\|} \sum_{b=1}^{\|d_j\|} SCP(w_i^r, w_j^b), \quad \text{with} \quad SCP(w_1, w_2) = \frac{P(w_1, w_2)^2}{P(w_1) \times P(w_2)} \tag{1}$$

[3] A centroid is represented by a p word feature vector $(w_k^1, w_k^2, w_k^3, \ldots, w_k^p)$.

Here, w_i^r (resp. w_j^b) corresponds to the token at the r^{th} (resp. b^{th}) position of the word feature vector d_i (resp. d_j). $\|d_i\|$ and $\|d_j\|$ respectively denote the total number of tokens in word feature vectors d_i and d_j. $SCP(w_i^r, w_j^b)$ is the Symmetric Conditional Probability (da Silva et al., 1999) where $P(.,.)$ is the joint probability of two tokens (w_1 and w_2) appearing in the same word feature vector and $P(.)$ is the marginal probability of any token appearing in a word feature vector.

Note that each cluster centroid is a word feature vector of varying number of tokens. Thus, Equation 2 is used to assign any data point (web snippet) d_j to a cluster t whose centroid has the maximum similarity value to d_j.

$$t = argmax_{k=1,...K} S(d_j, m_{\pi_k}) \tag{2}$$

K denotes the total number of clusters, d_j is the j^{th} web snippet, m_{π_k} is the centroid of the k^{th} cluster π_k and $S(d_j, m_{\pi_k})$ denotes similarity measurement between the point d_j and cluster centroid m_{π_k} defined in Equation 1.

4.3 Definition of Objective Functions

A string i represents a set of centroids to which web snippets can be assigned as seen in Section 4.2. As a consequence, each string i corresponds to a candidate partition of the data space. Now, in order to verify the domination of different solutions over other ones, objective functions must be defined. Compactness and separability are usually used in MOO clustering solutions. Here, compactness can be defined as the informational density of each cluster. This can be straightforwardly formulated as in Equation 3.

$$Compactness = \sum_{k=1}^{K} \sum_{d_i \in \pi_k} S(d_i, m_{\pi_k}) \tag{3}$$

Note that if tokens in a particular cluster are very similar to the cluster centroid then the corresponding $Compactness$ value would be maximized. Here our target is to form good clusters whose compactness in terms of similarity should be maximum.

The second objective function is cluster separability, which measures the dissimilarity between two cluster centroids. Indeed, the purpose of any clustering algorithm is to obtain compact similar typed clusters, which are dissimilar to each other. Here, we define separability as the minimization of the summation of similarities between each pair of cluster centroids. This is defined in Equation 4, where m_{π_k} and m_{π_o} are the centroids of clusters π_k and π_o, respectively.

$$Separability = \sum_{k=1}^{K} \sum_{o=k+1}^{K} S(m_{\pi_k}, m_{\pi_o}) \tag{4}$$

Finally, for a particular string, the following objectives $\{Compactness, \frac{1}{Separability}\}$ are maximized using the search capability of AMOSA.

4.4 Search Operators

In *MOO-clus*, AMOSA is used as the optimization strategy. For that purpose, three different types of mutation operations have been defined to suit the framework.

Mutation 1: This mutation operation is used to update the cluster center representation. Each token of cluster centroid is replaced by one token from the global vocabulary according to highest SCP similarity. This is applied individually to all tokens of a particular centroid if it is selected for mutation.

Mutation 2: This mutation operation is used to reduce the size of the string by 1. We randomly select a cluster centroid and thereafter all the tokens of this centroid are deleted from the string.

Mutation 3: This mutation is for increasing the size of string by 1 i.e. one new centroid is inserted in the string. For that purpose, we randomly choose p number of tokens from the global vocabulary and add it to the string.

Let be a string $< w_1\ w_2\ w_3\ w_4\ w_5\ w_6 >$ representing three cluster centroids (w_1, w_2), (w_3, w_4) and (w_5, w_6)[4]. For mutation 1, let position 2 be selected randomly. Each token of the word vector (w_3, w_4) will be changed by some token from the global vocabulary using SCP. Then, after change, the string will look like $< w_1\ w_2\ w_3^{new}\ w_4^{new}\ w_5\ w_6 >$. If mutation 2 is selected, a centroid will be removed from the string. Let centroid 3 be selected for deletion. The new string will look like $< w_1\ w_2\ w_3\ w_4 >$. In case of mutation 3, a new centroid will be added to the string. A new cluster centroid is generated choosing p=2 number of tokens from the global vocabulary. Let the randomly generated new cluster centroid to be added to the string be (w_7, w_8). After inclusion of this centroid, the string will be $< w_1\ w_2\ w_3\ w_4\ w_5\ w_6\ w_7\ w_8 >$. In our experiments, we have associated equal probability to each of these mutation operations. Thus, each mutation is applied in 33% cases of the cases.

5 Experimental Setup

5.1 Datasets

The main gold standards used for the evaluation of SRC algorithms are ODP-239 and MORESQUE[5]. In ODP-239 (Carpineto and Romano, 2010), each document is represented by a title and a web snippet and the subtopics are chosen from the top levels of DMOZ[6]. On the other hand, the subtopics in MORESQUE (Navigli and Crisafulli, 2010) follow a more natural distribution as they are defined based on the disambiguation pages of Wikipedia. As such, the subtopics cover most of the query-related senses. However, not all queries are Wikipedia related or ambiguous (e.g. "Olympic Games", which Wikipedia entry is not ambiguous, although there are many events related to this topic). As a consequence, it is clear that different results can be obtained from one data set to another. A quick summary of both data sets is presented in Table 1.

Dataset	# of queries	# of Subtopics Avg / Min / Max	# of Snippets
ODP-239	239	10 / 10 / 10	25580
MORESQUE	114	6.7 / 2 / 38	11402

Table 1: SRC gold standard data sets.

5.2 Evaluation Metrics

A successful SRC system must evidence high quality level clustering. Each query subtopic should ideally be represented by a unique cluster containing 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 this paper, we propose to use the F_{b^3}-measure (Amigó et al., 2009) to explore the Pareto front. In particular, F_{b^3} has been defined to evaluate cluster homogeneity, completeness, rag-bag and size-vs-quantity 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 in Equation 5

$$F_{b^3} = \frac{2 * P_{b^3} * R_{b^3}}{P_{b^3} + R_{b^3}}, \quad P_{b^3} = \frac{1}{N}\sum_{i=1}^{K}\sum_{d_j \in \pi_i}\frac{1}{|\pi_i|}\sum_{d_l \in \pi_i} g^*(d_j, d_l), \quad R_{b^3} = \frac{1}{N}\sum_{i=1}^{K}\sum_{d_j \in \pi_i^*}\frac{1}{|\pi_i^*|}\sum_{d_l \in \pi_i^*} g(d_j, d_l) \qquad (5)$$

where π_i is i^{th} cluster, π_i^* is the gold standard of the category i, and $g^*(.,.)$ and $g(.,.)$ are defined as follows:

$$g^*(d_i, d_j) = \begin{cases} 1 \Leftrightarrow \exists l: & d_i \in \pi_l^* \wedge d_j \in \pi_l^* \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad g(d_i, d_j) = \begin{cases} 1 \Leftrightarrow \exists l: & d_i \in \pi_l \wedge d_j \in \pi_l \\ 0 & \text{otherwise} \end{cases}.$$

[4] with p=2.

[5] AMBIENT has received less attention since the creation of ODP-239.

[6] http://www.dmoz.org [Last access: 14/03/2014].

Most SRC studies have also used the F_β-measure (F_β), which is defined in Equation 6.

$$F_\beta = \frac{(\beta^2+1)*P*R}{\beta^2*P+R}, \quad P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN} \tag{6}$$

where

$$TP = \sum_{i=1}^{K} \sum_{d_j \in \pi_i^*} \sum_{\substack{d_l \in \pi_i^* \\ l \neq j}} g(d_i, d_j), FP = \sum_{i=1}^{K} \sum_{d_j \in \pi_i} \sum_{\substack{d_l \in \pi_i \\ l \neq j}} (1 - g^*(d_i, d_j)), FN = \sum_{i=1}^{K} \sum_{d_j \in \pi_i^*} \sum_{\substack{d_l \in \pi_i^* \\ l \neq j}} (1 - g(d_i, d_j)).$$

6 Results and Discussion

In this evaluation, we used the open source framework GATE (Cunningham et al., 2013) without stopword removal for web snippet tokenization[7]. We executed *MOO-clus* over ODP-239 and MORESQUE. The parameters of *MOO-clus* are: $T_{min} = 0.01$, $T_{max} = 100$, $\alpha = 0.85$, $HL = 10$, $SL = 20$ and $iter = 15$. Note that, they have been determined after conducting a thorough sensitivity study. A first set of experiments have been conducted for different p values of tokens present in the centroid, namely in the range 2 to 5 in order to understand the behavior of *MOO-clus* w.r.t. centroid size[8]. Note that the partition with maximum F_{b^3} is choosen for each size of p[9]. Overall results are shown in Table 2.

	MORESQUE				ODP-239			
	MOO-clus				*MOO-clus*			
	2	3	4	5	2	3	4	5
$F_b{}^3$	0.477	0.491	0.497	**0.502**	0.478	0.481	**0.484**	0.481
F_1	0.661	0.666	**0.675**	0.658	0.379	0.379	**0.384**	0.381
F_2	0.750	**0.768**	0.764	0.742	0.534	0.536	**0.537**	0.535
F_5	0.831	**0.862**	0.846	0.820	0.717	**0.720**	0.716	0.715

Table 2: Evaluation results of *MOO-clus* over MORESQUE and ODP239 data sets.

Results show that for MORESQUE, *MOO-clus* obtains the highest F_{b^3} value for p=5. In particular, performance increases for higher values of p. For ODP-239, best results are reported for p=4, but evidence less sensitivity to the number of words in the centroids. Indeed, a marginal difference is obtained between all runs. In terms of F_β, the same behaviour is obtained for ODP-239. But, for MORESQUE, best results are provided for smaller values of p, namely p=3.

Two important comments must be pointed at. In the first place, F_{b^3} shows a steady behaviour compared to F_β when the data set changes. The conclusions drawn in (Amigó et al., 2009) reporting the superiority of F_{b^3} over F_β seem to be verified for the specific case of SRC. In the second place, *MOO-clus* evidences a marginal sensitivity to different p values. Indeed, for ODP-239, changing p between 2 and 5 words has a negligible impact on F_{b^3}. The figures show a different behaviour for MORESQUE but this can easily be explained. In MORESQUE, less queries are provided for test and the number of reference clusters varies between 2 and 38, with a majority of queries containing very few clusters (the average cluster size is 6.7). As such, small clustering errors may result in high deviations in the evaluation metrics. So, p can be seen as a non influent parameter for clustering purposes. In fact, increasing the value of p may exclusively allow a more descriptive power for cluster labeling.

We also compared *MOO-clus* to the current state-of-the-art text-based (endogenous) SRC algorithms: STC (Zamir and Etzioni, 1998), LINGO (Osinski and Weiss, 2005), OPTIMSRC (Carpineto and Romano, 2010), Bisecting Incremental K-means (BIK), GK-means (Moreno et al., 2013) and the combination STC-LINGO (Moreno and Dias, 2014). The results are illustrated in Table 3 where we provide values for all the metrics for open source implementations and reported values in the literature for the

[7]Note that keeping stop words is a challenging task as most methodologies withdraw these elements as they are hard to handle. This decision is supported by the fact that we aim to produce as much as possible language-independent solutions.

[8]Note that to ease the user effort in searching for information, the cluster label must be small and expressive. Typical configurations range between 3 to 5 to include multiword expressions.

[9]F_β metrics are calculated over the partition with highest F_{b^3} value.

other experiments i.e. OPTIMSRC, GK-means and STC-LINGO. In particular, the Min (resp. Max) column refers to the worst (resp. best) performance when varying p, the size of the centroid.

The results of Table 3 clearly show the performance improvements of our proposed methodology over existing text-based techniques for both data sets and most evaluation metrics. For ODP-239, *MOO-clus* attains the highest values with respect to F_1, F_2, F_5 and F_{b^3} metrics against all existing endogenous algorithms. For MORESQUE, our algorithm reaches highest performance over all state-of-the-art algorithms for F_1 and F_{b^3} metrics but marginally fails for F_2 and F_5 against GK-means.

		MOO-clus		*SOO SRC*				*Combination of SOO SRC*	
		Min	*Max*	GK-means	STC	LINGO	BIK	OPTIMSRC	STC-LINGO
MORESQUE	F_1	0.658	**0.675**	0.665	0.455	0.326	0.317	N/A	0.561
	F_2	0.742	0.768	**0.770**	0.392	0.260	0.269	N/A	N/A
	F_5	0.820	0.862	**0.872**	0.370	0.237	0.255	N/A	N/A
	${F_b}^3$	0.477	**0.502**	0.482	0.460	0.399	0.315	N/A	0.498
ODP-239	F_1	0.379	**0.384**	0.366	0.324	0.273	0.200	0.313	0.362
	F_2	0.534	**0.537**	0.416	0.319	0.167	0.173	0.341	N/A
	F_5	0.715	**0.720**	0.462	0.322	0.153	0.165	0.380	N/A
	${F_b}^3$	0.478	**0.484**	0.452	0.403	0.346	0.307	N/A	0.425

Table 3: Comparative results with respect to F_β and F_{b^3} metrics over the ODP-239 and MORESQUE datasets obtained by different SRC techniques.

It is important to notice that OPTIMSRC and STC-LINGO can be viewed as a combination of different SRC SOO solutions but still casted to a SOO solution. These previous results report interesting issues for SRC and confort the idea that the combination of different objective functions may lead to enhanced SRC algorithms. But, *MOO-clus* is capable to find better partitions than OPTIMSRC and STC-LINGO for all data sets and all evaluation metrics as reported in Table 3.

It is important to notice that the *MOO-clus* provides a set of partitions with automatic definition of the number of clusters. So, defining one unique solution is an important issue for SRC. So far, we have provided results for the best partition evaluated by F_{b^3}. However, deeper analysis of all the partitions on the Pareto front must be endeavoured. Results are reported for F_{b^3} only as all other metrics behave correspondingly and are reported in Table 4.

	MORESQUE				ODP-239			
	2	3	4	5	2	3	4	5
Min	0.428	0.464	0.464	0.462	0.396	0.401	0.403	0.408
Max	0.477	0.491	0.497	0.502	0.478	0.481	0.484	0.481
Avg.	0.454	0.479	0.482	0.486	0.443	0.447	0.448	0.449

Table 4: F_{b^3} evaluation results of the Pareto front.

Figures show the validity of each individual solution of the Pareto front. In the worst case, *MOO-clus* produces similar results compared to the hard baseline STC. On average, it reaches the results of GK-means and the highest performance values can be found on the Pareto front. The correct identification of the best partition is still an open issue and can be compared to the automatic selection of K clusters, which is a hard task as shown in recent studies (Scaiella et al., 2012; Marco and Navigli, 2013).

7 Conclusions

In this paper, we proposed the first attempt[10] to define the SRC task as a multi-objective problem. For that purpose, we defined two objective functions, which are simultaneously optimized through the archived multi-objective simulated annealing framework called AMOSA. A correct definition of the task allowed to take advantage of the most recent advances in terms of endogenous SRC algorithms as well as the most powerful techniques for multi-objective clustering. The performance of *MOO-clus* has been evaluated over two gold standard data sets, ODP-239 and MORESQUE for different evaluation metrics, F_1 and F_{b^3}.

[10] As far as we know.

Results showed that our proposal steadily outperforms all existing state-of-the-art text-based endogenous SRC algorithms and approaches recent knowledge-driven exogenous strategies (Scaiella et al., 2012), which reach F_1=0.413 for ODP-239[11].

As future works, we propose to use MOO clustering in a strict meta learning way, where any labeled-based SOO solution is defined by specific $Compactness$ and $Separability$ functions. Another research direction is the definition of the Dual representation proposed by (Moreno et al., 2014) as a MOO problem. Finally, new objective functions can be defined to measure the quality of the labels, which may integrate meaningful multiword expressions or named entities.

Acknowledgement

We would like to thank the CNRS to provide Sriparna Saha with a 6 months internship at the GREYC Laboratory of the Normandie University.

References

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 (SIGIR)*, pages 643–652.

Sanghamitra Bandyopadhyay, Anirban Mukhopadhyay, and Ujjwal Maulik. 2007. An improved algorithm for clustering gene expression data. *Bioinformatics*, 23(21):2859–2865.

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

James C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

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

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

Claudio Carpineto, Stanislaw Osinski, Giovanni Romano, and Dawid Weiss. 2009. A survey of web clustering engines. *ACM Computing Surveys*, 41(3):1–38.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854.

Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira 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.

Kalyanmoy Deb. 2009. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley.

Julia Handl and Joshua Knowles. 2007. An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, 11:56–76.

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

Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Anirban Mukhopadhyay. 2011. *Multiobjective Genetic Algorithms for Clustering - Applications in Data Mining and Bioinformatics*. Springer.

José G. Moreno and Gaël Dias. 2014. Easy web search results clustering: When baselines can reach state-of-the-art algorithms. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1–5.

[11]Note that results of (Marco and Navigli, 2013) are not reported in this paper as the authors do not use the standard versions of MORESQUE and do not provide experiments for ODP-239.

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

José G. Moreno, Gaël Dias, and Guillaume Cleuziou. 2014. Query log driven web search results clustering. In *Proceedings of the 37th Annual ACM SIGIR Conference (SIGIR)*.

Katharina Morik, Andreas Kaspari, Michael Wurst, and Marcin Skirzynsk. 2012. Multi-objective frequent termset clustering. *Knowledge Information Systems*, 30(3):715–738.

Anirban Mukhopadhyay and Ujjwal Maulik. 2009. Unsupervised pixel classification in satellite imagery using multiobjective fuzzy clustering combined with SVM classifier. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1132–1138.

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 (EMNLP)*, pages 116–126.

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

Sriparna Saha and Sanghamitra Bandyopadhyay. 2010. A symmetry based multiobjective clustering technique for automatic evolution of clusters. *Pattern Recognition*, 43(3):738–751.

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

Xuanli L. Xie and Gerardo Beni. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:841–847.

Oren Zamir and Oren 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.