Query Log Driven Web Search Results Clustering

Jose G. Moreno
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
jose.moreno@unicaen.fr

Gaël Dias
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
gael.dias@unicaen.fr

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

ABSTRACT

Different important studies in Web search results clustering have recently shown increasing performances motivated by the use of external resources. Following this trend, we present a new algorithm called *Dual C-Means*, which provides a theoretical background for clustering in different representation spaces. Its originality relies on the fact that external resources can drive the clustering process as well as the labeling task in a single step. To validate our hypotheses, a series of experiments are conducted over different standard datasets and in particular over a new dataset built from the TREC Web Track 2012 to take into account query logs information. The comprehensive empirical evaluation of the proposed approach demonstrates its significant advantages over traditional clustering and labeling techniques.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*clustering*

General Terms

Algorithms, Experimentation

Keywords

Web Search Results Clustering, Dual *C*-means, Automatic Labeling, Evaluation

1. INTRODUCTION

Web search results clustering (SRC), also known as post-retrieval clustering, multifaceted clustering or ephemeral clustering has received much attention for the past twenty years. SRC systems return meaningful labeled clusters from a set of Web snippets retrieved from any Web search engine for a given user's query. So far, most works have focused on the study of topical clustering [9] although some studies have been appearing in temporal clustering [1] and geospatial clustering [39]. As a consequence, SRC systems can be particularly useful to understand query intents (topical clustering) and query diversity (temporal/geospatial clustering). In this paper, we particularly focus on topical SRC.

As opposed to classical text clustering, SRC must deal with small text fragments (Web snippets) and be processed in run-time. As a consequence, it is hard to implement efficiently and effectively [9]. So, most successful methodologies follow a monothetic approach [40, 10, 12, 35]. The underlying idea is to discover the most discriminant topical words in the collection and group together Web snippets containing these relevant terms. On the other hand, the polythetic approach, in which the main idea is to represent Web snippets based on the Vector Space Model (VSM) has received less attention [18, 22, 41, 30]. The main reason is the fact that the labeling process is a surprisingly hard extra task [9].

Our research is motivated by the fact that the adequate combination of the polythetic and monothetic approaches in a single algorithm should lead to improved performance over three important factors in SRC: *clustering accuracy*, *labeling quality* and *partitionning shape*. For that purpose, we present a new algorithm called *Dual C-Means*, which provides a theoretical background for dual-representation clustering. Its originality relies on the fact that different representation spaces can drive the clustering process as well as the labeling task in a single step.

We evaluated the proposed algorithm over different metrics (e.g. F_1^N [13], F_{b^3} [2], ARI [37], *D#-nDCG* [34]), well-studied datasets (e.g. ODP-239 [10], SEMEVAL [28]) and different representation spaces (e.g. text and query logs). The results show that the combination of the VSM representation of Web snippets and a query-log-based representation of cluster centroids achieves the best configuration for the SRC task. In particular, increased performance is shown against most SRC solutions (e.g. STC [40], LINGO [30], TOPICAL [35], LDA [7]). Our main contributions are :

- A new algorithm (Dual *C*-Means), which can be seen as an extension of *K*-means [21] for dual-representation spaces;
- An instantiation of the Dual *C*-Means for SRC, which takes advantage of external resources such as query logs to improve clustering accuracy, labeling quality and partitioning shape;
- A new annotated dataset (WEBSRC401) based on the TREC Web Track 2012 for full SRC evaluation over the Web.

In the next section, we present the most important recent studies for SRC. In the third section, we present the general model of the Dual *C*-Means algorithm and its instantiation in the context of SRC. In the fourth section, we explain the construction of the WEBSRC401 dataset. In the fifth and sixth sections, we present the experimental setups and show the results obatined for different strategies over an exhaustive set of well-known evaluation metrics,

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609583.

datasets and state-of-the-art algorithms. Finally, we draw some conclusions about our experiments and propose new perspectives.

2. RELATED WORK

A good survey of SRC methodologies can be found in [9]. As a consequence, we give a brief overview of older methodologies and focus on more recent works. The first important work in SRC is certainly proposed by [18]. They define a polythetic approach based on the VSM representation where similarity between documents is computed with cosine similarity measure. Then, a non-hierarchical partitioning strategy called fractionation is performed to discover the number of clusters suggested by the user. Initial results show that their "*approach to document clustering is one which can produce significant improvements over similarity search ranking alone*". Although they present the foundations of SRC, labeling is not tackled and evaluation is based on a small dataset and a limited user study.

In order to propose a more realistic solution, which includes labeling, [40] defined the Suffix Tree Clustering (STC) algorithm. They propose a monothetic clustering technique, which merges base clusters with high string overlap. Instead of using the VSM representation, they propose to represent Web snippets as compact tries. Their evaluation over a small set of 10 queries shows that STC outperforms group-average agglomerative hierarchical clustering, K-Means, buckshot, fractionation and single-pass algorithms. STC is still considered as a hard baseline to compete with.

Later, [30] proposed a polythetic approach called LINGO, which takes into account the string representation proposed by [40]. They first extract frequent phrases based on suffix-arrays. Then, they reduce the term-document matrix (defined as a VSM) using Single Value Decomposition to discover latent structures. Finally, they match group descriptions with the extracted topics and assign relevant documents to them. LINGO is evaluated with 7 users over a set of 4 search results and as such, no conclusive remarks can be drawn. However, their publicly available implementations of LINGO, STC and BiKM (Bi-section K-means) provide researchers with useful tools to build SRC systems.

More recently, [10] showed that the characteristics of the outputs of SRC algorithms suggest the adoption of a meta clustering approach. For that purpose, they introduce a novel criterion to measure the concordance of two partitions of Web snippets into different clusters based on the information content associated with the decisions made by the partitions on single pairs of Web snippets. A meta clustering phase is then casted to an optimization problem of the concordance between the clustering combination and the given set of clusterings. The results of their OPTIMSRC system demonstrate that meta clustering is superior over individual clustering techniques. In particular, they propose a dataset called ODP-239, which is widely used in the community.

Another polythetic methodology is proposed in [26]. Their underlying idea is that informative text similarity measures can improve SRC by adequately capturing the latent semantics conveyed by Web snippets. They propose a K-means based algorithm called GK-means within which a new objective function defined for a third-order similarity measure must be maximized. As different partitions are possible depending on the K value, they propose an automatic stopping criterion to retrieve one "optimal" clustering solution. Their main contribution is the fact that labels are built during the clustering process thus avoiding an extra processing step. Their results show improvements for ODP-239 in terms of F_{b^3} over all text-based SRC algorithms.

While all studies mentioned so far treat the task of SRC as a text-based problem, some other works propose to introduce external resources. The first relevant work is presented in [16] where Web snippets are enriched with anchor text information and high quality indexes extracted from DMOZ. The underlying idea of their monothetic approach called SNAKET is that better labeling and clustering can be obtained from these external resources. Results over a non-standard dataset show that the introduction of external information improves Precision at different clustering levels. [16] certainly proposed a new trend in SRC.

Following the same idea, [35] proposed TOPICAL, a top performing SRC system over ODP-239 dataset. They propose to move away from the bag of words representation towards a graph of topics paradigm derived from TAGME, a wikification algorithm [38]. Each Web snippet is annotated with a set of topics, which are represented by Wikipedia articles. A bipartite-like graph structure is built where nodes are either Web snippets or topics and edges are either topic-to-topic or topic-to-snippet. Then, a spectral-like clustering algorithm is run over the graph to discover relevant clusters and meaningful labels. TOPICAL is an interesting approach as clustering is driven by the presence of Wikipedia titles in Web snippets and inderectly assures the quality of the labeling.

Another idea has recently been proposed in [13], which relies on Web n-grams. In order to better capture the similarity between Web snippets, a first step consists in building a co-occurence graph based on Dice coefficient calculated over the Google Web1T corpus [8] from which senses are discovered by word sense induction algorithms. Each Web snippet is represented as a bag-of-words (polythetic approach) but Similarity is computed over discovered word senses. Their experiments show that enhanced diversification and clustering performance results can be obtained based on the adjusted RandIndex [37] for a specific dataset built for ambiguous queries (MORESQUE). Recently, researchers from the same team proposed a new dataset within the context of the SEMEVAL task 11 [28], in which the goal is to provide an evaluation framework for the objective comparison of word sense disambiguation and induction algorithms in SRC for ambiguous queries.

All works propose interesting issues for SRC. On one hand, the monothetic approach mainly focuses on the identification of strong meaningful labels. The underlying idea is that good labels are a key factor for the success of user experience in Web search. On the other hand, the polythetic approach concentrates on discovering high quality clusters and the labeling task is usually treated as a separate process. The subjacent motivation is that good clustering should be provided to improve user experience in search for information. Moreover, recent studies show that the introduction of external resources improves overall results.

In this paper, we propose that both monothetic and polythetic approaches should be combined in a single algorithm capable of accepting external resources. For that purpose, we present the Dual C-Means algorithm, which extends the well-known K-Means for dual representation spaces. It particularity relies on the fact that different representation spaces compete to reach high clustering quality and meaningful labeling. In particular, we propose that query logs are introduced as external information to ensure quality labeling and drive the clustering process. The main characteritics of our proposal are as follows:

- New combination of polythetic and monothetic approaches in one single algorithm;
- Introduction of dual representations for Web snippets allowing the introduction of external resources;
- Theoretical framework based on an extension of K-means;
- First proposal with query logs as external resource for SRC.

3. DUAL C-MEANS ALGORITHM

This section is devoted to the presentation of the Dual *C*-Means algorithm that extends the classical *K*-means [21] for dual representation spaces. In the first subsection, we present the general model and in the second one we propose its instantiation for the specific task of SRC.

3.1 General Model

Let S be a dataset to partition where each data $s_i \in S$ is described on a representation space E_1 and additionally, E_2 denotes another space supporting cluster representation. We hypothesize the existence of a function $d: E_1 \times E_2 \rightarrow \mathbb{R}^+$ quantifying the dissimilarity between any data from E_1 and any cluster representative (cluster centroid) from E_2. The new proposed clustering model (Dual *C*-Means) is driven by the objective criterion defined in Equation 1, which must be minimized.

$$J_{dcm}(\Pi, M) = \sum_{k=1}^{c} \sum_{s_i \in \pi_k} d(s_i, m_k) \quad (1)$$

As illustrated in Figure 1, the aim of the minimization of $J_{dcm}(\Pi, M)$ is to find a partition of S into c clusters ($\Pi = \{\pi_1, \ldots, \pi_c\}$) such that in each cluster π_k any object is as closed as possible to a common cluster representative m_k ($M = \{m_1, \ldots, m_c\}$).

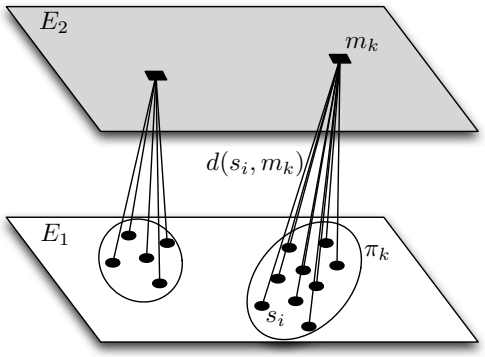

Figure 1: Dual *C*-Means aims to discover clusters of objects in E_1 closed to a common cluster representative in E_2.

The optimization process can be achieved by an usual dynamic reallocation algorithm starting with a random initial clustering Π^0 and then iterating the following two steps (Update and Assignment) until convergence:

1. Update: compute new optimal cluster representatives M^{t+1} considering a fixed partition Π^t,
2. Assignment: compute new optimal assignments Π^{t+1} considering fixed cluster representatives M^{t+1} and use the following rule to assign each object to its closest representative:

$$\forall s_i,\ s_i \in \pi_k \Leftrightarrow k = \arg\min_{l=1,\ldots,c} d(s_i, m_l).$$

Note that the update of cluster representatives has to be defined depending on both the dissimilarity measure $d(.,.)$ and the representative space E_2 in order to ensure that the objective criterion $J_{dcm}(.,.)$ decreases. Let us also notice that in the specific case where $E_1 = E_2 = \mathbb{R}^n$ and the squared euclidean distance is chosen as dissimilarity $d(.,.)$, the Dual *C*-Means algorithm comes down exactly to the usual *K*-means algorithm ($m_k^{t+1} = \sum_{s_i \in \pi_k^t} s_i / |\pi_k^t|$). Finally, such as *K*-means, Dual *C*-Means is sensitive to random initialization and requires the number of expected clusters (C) as parameter[1].

3.2 Instantiation in the SRC Context

In the context of SRC, objects are naturally Web snippets represented in the E_1 space ($s_i \in S$) and cluster representatives are labels represented in the E_2 space ($m_k \in M$).

The crucial hypothesis of the Dual *C*-Means algorithm is the existence of a dissimilarity metric $d(.,.)$ capable of comparing objects from different feature spaces. For that purpose, a matching process between the two feature sets is required that can be formalized as a transition matrix P ($p_1 \times p_2$) quantifying this matching for each of the p_1 features defined in E_1 with each of the p_2 features from E_2.

Without loss of generality, we define a generic dissimilarity measure considering such a transition matrix in Equation 2 where m_k^T is the transposed label vector, $s_i P m_k^T$ quantifies a similarity between a Web snippet s_i and a label m_k, and α is a constant to adjust in order to ensure dissimilarity values in $\mathbb{R}^+$.

$$d(s_i, m_k) = \alpha - s_i P m_k^T \quad (2)$$

Such a dissimilarity form allows us to rewrite the Dual *C*-Means algorithm as a maximization problem defined in Equation 3.

$$\min_{\Pi, M} \sum_{k=1}^{c} \sum_{s_i \in \pi_k} d(s_i, m_k) \Leftrightarrow \max_{\Pi, M} \sum_{k=1}^{c} \sum_{s_i \in \pi_k} s_i P m_k^T \quad (3)$$

Let us notice that when the label space E_2 is unconstrained (e.g. $E_2 = \mathbb{R}^{p_2}$), the resolution of Equation 3 has no sense ($M = +\infty$). But in the SRC context, a small set of words (i.e. the labels) are usually chosen to help the user in his search for information. Thus, we consider two vocabularies V_1 and V_2 defining the two feature spaces E_1 and E_2 respectively. We constrain Web snippet descriptions to be word distributions over V_1 ($s_{i,j} \in [0,1]\ \forall i,j$ and $\sum_{j=1}^{p_1} s_{i,j} = 1$) and cluster labels to subsets of p words from V_2 ($E_2 = \{m_k \in \{0,1\}^{p_2} | \sum_{l=1}^{p_2} m_{k,l} = p\}$).

Within that context, the computation of optimal cluster labels is a discrete optimization process solved for each cluster π_k independently, by first sorting the vocabulary V_2 from the most relevant word (l_1^k) to the less relevant one ($l_{p_2}^k$) using the relevance function defined in Equation 4

$$\forall l, k\ \ \lambda_k(l) = \sum_{s_i \in \pi_k} s_i P_{.,l} \quad (4)$$

and then defining a cluster label vector m_k as the combination of the p most relevant words from V_2 for the snippets in π_k as proposed in Equation 5.

$$m_{k,l} = \begin{cases} 1 \text{ if } l \geq l_p^k \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

It is interesting to notice that the *GK*-means, recently proposed by [26], falls into such an SRC instantiation of the Dual *C*-Means algorithm if the following constraints are true:

- Web snippet and label representation spaces are not dissociated (i.e. $V_1 = V_2$) thus not taking benefit from the duality of the clustering algorithm;

[1]These issues will be tackled in the Evaluation section.

- The transition matrix P is computed with the Symmetric Conditional Probability (SCP [36]) or the Pointwise Mutual Information (PMI [11]) on the unique vocabulary $V_1 = V_2$.

To make use of the duality concept from the new proposed algorithm in the SRC context, we suggest differentiating the two vocabularies V_1 and V_2. First, V_1 is defined as the bag of words occurring in all Web snippets retrieved for a given query. Second, if we consider a set Y of query logs, the vocabulary V_2 is defined by the bag of words occurring in Y and E_2 is restricted to the set of query logs defined as distributions in the vector space model induced by V_2. This situation is formalized in Equation 6 with β_i denoting the size of the query log y_i.

$$E_2 = \{y_i \in \{0, \frac{1}{\beta_i}\}^{p_2} | \sum_{j=1}^{p_2} y_{i,j} = 1 \text{ and } y_i \in Y\} \quad (6)$$

As such clustering is polythetic but query log driven. Figure 2 illustrates the instantiation of the Dual C-Means algorithm in the SRC context where the restricted set of available query logs guides the cluster formation process.

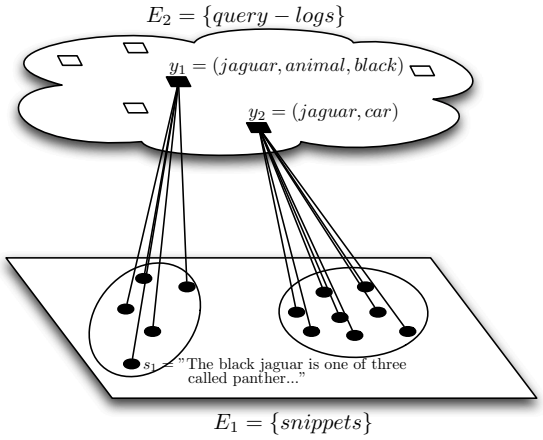

Figure 2: Example of the Dual C-Means instantiated for the SRC context with query logs as cluster label space.

4. THE WEBSRC401 DATASET

Different gold standards have been used for the evaluation of SRC algorithms among which the most cited are: AMBIENT [6], ODP-239 [10], MORESQUE [27] and SEMEVAL [28]. As ODP-239 is an evolution of AMBIENT and SEMEVAL is the next generation of MORESQUE, we will only give an overview of the most recent datasets.

In ODP-239, each document is represented by a title and a Web snippet and the subtopics are chosen from the top levels of DMOZ[2]. However, this dataset does not represent the typical kind of results obtained through querying a given search engine as the number of possible subtopics is always equal to 10. It is clear that this structure clearly differs from a typical Web results set. Moreover, queries are not extracted from query logs but rather chosen based on the categories present in DMOZ. However, it is a publicly available dataset that allows us to conduct experiments to evaluate clustering accuracy.

[2]http://www.dmoz.org [Last access: 27/01/2014].

On the other hand, the subtopics in SEMEVAL follow a more natural distribution as they are defined based on the disambiguation pages of Wikipedia. As such, these subtopics are likely to cover most of the senses present in the Web for the 100 evaluated queries. However, SEMEVAL is built to specifically deal with ambiguous queries, which are self-contained in Wikipedia. But, it is clear that not all queries in general are Wikipedia articles or ambiguous. For example, many queries are multifaceted but not ambiguous [19]. Let us take "Olympic Games". Its Wikipedia entry is not ambiguous but it presents many different facets such as History, Logos, Year dates or Cities, to name but a few.

As a consequence, it is clear that different results can be obtained from one dataset to another. A quick summary of both datasets is presented in Table 1.

Dataset	# of queries	# of Subtopics Avg / Min / Max	# of Web snippets
ODP-239	239	10 / 10 / 10	25580
SEMEVAL	100	7.7 / 2 / 19	6400
WEBSRC401	50	3.9 / 3 / 6	5560

Table 1: Description of the SRC gold standard datasets.

To afford a more realistic situation in the context of Web search results, we propose a new SRC dataset based on the ClueWeb09 Category B text collection (CCB)[3], which comprises about 50 million English-language pages, including the entirety of the English-language Wikipedia and task descriptions of the TREC Web Track 2012. The goal of TREC Web Track 2012 is to return a ranked list of Web pages that together provide complete topical coverage of a given query, while avoiding excessive redundancy of the subtopics in the result list. In particular, each topic contains a query field, a description field and several subtopic fields which can be ambiguous or multifaceted. And for each topic, a judgement file (i.e. qrel) includes the list of relevant Web pages from CCB and the manually attributed grade of the Web page subtopic.

Instead of retrieving relevant Web pages, we are interested in obtaining relevant clusters (i.e. Web pages with the same subtopic) with high coverage of all the subtopics. So, we propose transforming the data available in the TREC Web Track 2012 in a typical SRC format [10], which result in the WEBSRC401 dataset[4]. First, for each Web page considered as query-relevant, its Web snippet is retrieved using the *SnippetGenerator* function of ChatNoir[5]. By default, a Web snippet composed of a maximum of 500 characters found around the query words is provided.

Secondly, for each query, its subtopics are defined as in the TREC Web Track 2012 and each qrel is encoded in a new format, which contains the Web page id, the subtopic id and the query[6]. Additionally, it is important to notice that the WEBSRC401 dataset facilitates the evaluation of new techniques based on more complex resources provided by researchers as it is based on the well-studied ClueWeb09. For example, cluster ranking or spam cluster filtering studies could be endeavored with the PageRank scores and the spam rankings of ClueWeb09 dataset which are publicly available.

5. CLUSTERING EVALUATION

As mentioned in [9], evaluating SRC systems is a hard task. Indeed, the evaluation process is three-fold. A successful SRC sys-

[3]http://lemurproject.org/clueweb09/ [Last access: 27/01/2014]
[4]http://websrc401.greyc.fr/ [Last access: 10/05/2014].
[5]http://chatnoir.webis.de/ [Last access: 27/01/2014].
[6]Note that these steps could be used to extend the dataset with the TREC Web tracks of the years 2009, 2010 and 2011.

tem must discover relevant topical clusters (*clustering accuracy*) and propose meaningful labels at the same time (*labeling quality*). We will also see in our experiments that *partition shape* is also an important factor to study.

5.1 Evaluation of SRC

Firstly, a successful SRC system must evidence high quality level clustering. Ideally, each query subtopic should be represented by a unique cluster containing all the relevant Web pages inside. However, this task is far from being achievable. As such, this constraint can be reformulated as for the TREC Web Track 2012: the task of SRC systems is to provide complete topical cluster coverage of a given query, while avoiding excessive redundancy of the subtopics in the result list of clusters.

Secondly, SRC systems should present meaningful labels to the user to ease their search for information. As such, the evaluation of the labeling task is of the utmost importance. As far as we know, only [10, 35] propose the evaluation of both dimensions. However, their experiments are not reproducible as they rely on manually annotated datasets, which are not publicly available.

Thirdly, SRC differs from classical text clustering as the partitioning shape, more precisely the distribution of the Web snippets into clusters, shows evidence of some particularity. Indeed, it is well-known that subtopics on the Web are not equally distributed. For example, for the query "Apple", it is much more likely to find Web snippets related to the company than the concept of fruit. In particular, we will see in our experiments that not all evaluation metrics cover this situation.

In the next sections, we propose a complete set of repeatable experiments to give an exhaustive overview of the SRC field. We start by focusing on the experimental setups.

5.2 Experimental Setups

In this section, we propose the comparison of different configurations of the Dual *C*-Means to several state-of-the-art algorithms using well-studied evaluation metrics.

Dual C-Means Configurations.

The originality of the Dual *C*-Means is to embody a great number of possible configurations due to the expressiveness of its model. In this paper, we will particularly focus on two main issues. The *first one* deals with using different similarity measures to compute the transition matrix P. The underlying idea is supported by the fact that different word similarity measures produce different results [31]. As a consequence, we aim to understand their impact on the SRC task. The *second one* aims to test our initial hypothesis stating that the introduction of external resources can improve SRC. As a consequence, we propose two different space representations: text-text (i.e. $V_1 = V_2$) and text-query logs (i.e. $V_1 \neq V_2$).

Word Similarity Measures.

The use of word similarity metrics is an important and interchangeable component of our algorithm encoded in the transition matrix. In this study, we propose the comparison of a total of five collocation metrics[7]. In particular, we used the Symmetric Conditional Probability (SCP) [36], the Pointwise Mutual Information (PMI) [11], the Dice coefficient [14], the LogLikelihood ratio (LogLike) [15] and Φ^2 [17]. Each metric is defined in Table 2. The expressiveness of the Dual *C*-means permits the definition of different types of word similarity measures. As a consequence, we also compute word-word similarity based on the VSM representation. In particular, for each snippet $s_i \in S$, a simple word-word similarity measure is $S^T S$ where S^T is the transposed of the snippet-term matrix S. In this case, $P = S^T S$. Another interesting similarity measure is LSA [20], which can be formulated as follows: $P = U\ \Lambda_e\ U^T$ where $U\ \Lambda\ U^T$ is the eigen decomposition of $S^T S$, and e is the number of highest eigen values selected to represent the latent space[8].

SRC Algorithms.

We aim to compare our algorithm to the most competitive strategies proposed so far in the SRC literature. For that purpose, we show the results of STC [40], LINGO [30], TOPICAL [35] and LDA [7]. It is worth noticing that for evaluation purposes, we developed an open source implementation[9] of TOPICAL using the Wikipedia Miner API proposed by [24] and the spectral algorithm proposed by [29] included in SCIKit learning tool[10]. For LINGO and STC algorithms, we used the Carrot2 API[11]. And for LDA, we used the topic modeling package included in MALLET toolkit [23]. The parameters were set following the toolkit instructions (i.e. stop-words removal, $\alpha_t = 0.01$, $\beta_w = 0.01$ and limited to 1000 iterations) and the cluster membership is assigned taken the maximum topic probability value.

Evaluation Metrics.

Different metrics have been proposed to evaluate text clustering. Within this paper, we present the results for the most relevant metrics. The first complete study in terms of evaluation has certainly been proposed by [10]. In the specific case of SRC, the authors propose the F_1^C metric, which is a specific implementation of the more general F_β measure. Other metrics have also been proposed. For example, the F_{b^3} measure [2] addresses many important problems in clustering such as cluster homogeneity, completeness, rag-bag and size-vs-quantity constraints, and has shown interesting properties for the SRC task as formulated in [26]. Two other important metrics have been studied in [13]: F_1^N and the Adjusted RandIndex (ARI) [37]. In particular, F_1^N can be seen as a complementary metric of F_1^C as it is also based on the classical F_β measure but computed in a different manner[12], while ARI evidences an interesting property for SRC. While it measures clustering accuracy, it also takes into account the fact that a given partition shows a similar partitioning shape compared with the reference gold standard. The underlying idea is that the number of clusters and the average number of Web snippets in each cluster approximate as much as possible the reference clustering. An illustration of this situation can be seen in [25] although the authors do not refer to this issue as an important one for SRC. In terms of implementation, we used the Java evaluator[13] to compute both F_1^N and ARI evaluation metrics, and the implementation provided by [3][14] to compute F_{b^3}. In Table 3, we defined all the metrics used for our experiments.

[7] It is clear that a great deal of association measures that could be tested exist. However, we selected the ones which best complement themselves.

[8] In our experiments, this value was set to the minimum which guarantees that $\sum_{i=1}^{e} \Lambda_i \geq 0.9 \sum_{i=1}^{p1} \Lambda_i$.

[9] This implementation is publicly available upon request.

[10] http://scikit-learn.org/stable/ [Last access: 27/01/2014].

[11] http://search.carrot2.org/stable/search [Last access: 27/01/2014].

[12] Let us notice that these are two F_1 measures, which computation is defined differently in [10] and [13].

[13] http://www.cs.york.ac.uk/semeval-2013/task11/index.php?id=data [Last access: 27/01/2014].

[14] http://nlp.uned.es/~enrique/software/RS.zip [Last access: 27/01/2014].

Collocation Metric	Formula
$SCP(w_i,w_j)$	$\frac{P(w_i,w_j)^2}{P(w_i)*P(w_j)}$
$PMI(w_i,w_j)$	$\log_2 \frac{P(w_i,w_j)}{P(w_i)*P(w_j)}$
$DICE(w_i,w_j)$	$\frac{2*f(w_i,w_j)}{f(w_i)+f(w_j)}$
$LogLike(w_i,w_j)$	$-2*logLike(f(w_i,w_j),f(w_i),\frac{f(w_j)}{N})+logLike(f(w_j)-f(w_i,w_j),N-f(w_i),\frac{f(w_j)}{N})$ $-logLike(f(w_i,w_j),f(w_i),\frac{f(w_i,w_j)}{f(w_i)})-logLike(f(w_j)-f(w_i,w_j),N-f(w_i),\frac{f(w_j)-f(w_i,w_j)}{N-f(w_i))}$ where $logLike(a,b,c)=(a*Log(c))+((b-a)*Log(1-c))$
$\Phi^2(w_i,w_j)$	$\frac{P(w_i,w_j)-P(w_i)*P(w_j)}{P(w_i)*P(w_j)*(1-P(w_i))*(1-P(w_j))}$

Table 2: Collocation metrics used in our framework where $P(w_i,w_j)$ is the joint probability of words w_i and w_j, $P(w_i)$ is the marginal probability of the word w_i, $f(w_i,w_j)$ is the frequency of word pairs (w_i,w_j), $f(w_i)$ is the frequency of the word w_i and N is the number of retrieved Web snippets.

Evaluation Metric	where
$F_1^C=\frac{2*P*R}{P+R}$	$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN},TP=\sum_{z=1}^{k^*}\sum_{x_j\in\pi_z^*}\sum_{x_l\in\pi_z^*,l\neq j}g_0(x_j,x_l),FP=\sum_{i=1}^{k}\sum_{x_j\in\pi_i}\sum_{x_l\in\pi_i,l\neq j}(1-g_0^*(x_j,x_l)),$ $FN=\sum_{z=1}^{k^*}\sum_{x_j\in\pi_z^*}\sum_{x_l\in\pi_z^*,l\neq j}(1-g_0(x_j,x_l))$
$F_{b^3}=\frac{2*P_{b^3}*R_{b^3}}{P_{b^3}+R_{b^3}}$	$P_{b^3}=\frac{1}{N}\sum_{i=1}^{k}\frac{1}{\lvert\pi_i\rvert}\sum_{x_j\in\pi_i}\sum_{x_l\in\pi_i}g_0^*(x_j,x_l),R_{b^3}=\frac{1}{N}\sum_{z=1}^{k^*}\frac{1}{\lvert\pi_z^*\rvert}\sum_{x_j\in\pi_z^*}\sum_{x_l\in\pi_z^*}g_0(x_j,x_l)$
$F_1^N=\frac{2*P*R}{P+R}$	$P=\frac{1}{\sum_{i=1}^{k}\lvert\pi_i\rvert}\sum_{i=1}^{k}\max_{\pi_z^*}(f(\pi_z^*,\pi_i)),R=\frac{1}{\sum_{z=1}^{k^*}\lvert\pi_z^*\rvert}\sum_{z=1}^{k^*}f(\pi_z^*,\bigcup_{\pi_i\in\Pi_z}\pi_i)$ $\pi_b\in\Pi_z\iff z=\arg\max_a(f(\pi_a^*,\pi_b)),f(\pi_a^*,\pi_b)=\sum_{x_j\in\pi_a^*}\sum_{x_l\in\pi_b}g_1(x_j,x_l)$
$ARI(\Pi,\Pi^*)$	$ARI(\Pi,\Pi^*)=\frac{RI(\Pi,\Pi^*)-E(RI(\Pi,\Pi^*))}{maxRI(\Pi,\Pi^*)-E(RI(\Pi,\Pi^*))}$ where $RI(\Pi,\Pi^*)=\frac{TP+TN}{TP+FP+FN+TN},TN=N-TP-FP-FN$
and	$g_0(x_i,x_j)=\begin{cases}1\iff\exists l:x_i\in\pi_l\wedge x_j\in\pi_l\\0,otherwise\end{cases}$ $g_0^*(x_i,x_j)=\begin{cases}1\iff\exists l:\quad x_i\in\pi_l^*\wedge x_j\in\pi_l^*\\0,otherwise\end{cases}$ where π_i is the cluster solution i ($\Pi=\cup\pi_i$) and π_i^* is the gold standard of the category i ($\Pi^*=\cup\pi_i^*$). $g_1(x_i,x_j)=\begin{cases}1\iff x_i=x_j\\0,otherwise\end{cases}$

Table 3: Clustering Evaluation Metrics.

Text Processing and Implementation.

All Web snippets were tokenized with the GATE platform[15] but we did not apply stop-words removal so that we can propose a language-independent solution. In terms of dynamic reallocation algorithm, we used the optimized version of *K*-means++ proposed in [4] as the intialization process is semi-deterministic[16] and there exists an efficient implementation called Scalable K-means++ [5].

5.3 Clustering Results

A great deal of experiments have been performed to achieve conclusive results. We first propose evaluating the clustering accuracy of the Dual *C*-Means against different state-of-the-art algorithms. For that purpose, we propose an exhaustive search as in [35], whose underlying idea is to evaluate the behavior of a given algorithm along with the increasing number of output partitions. In this first set of experiments, we pretend to understand the clustering quality of our approach when only text information is taken into account (i.e. $V_1=V_2$ and the number of p words composing the centroids is set to 2) and compare it to state-of-the-art algorithms. In particular, we present the results for 20 runs ($K=2..20$) and illustrate the F_{b^3} values over ODP-239 and WEBSRC401 in Figure 3. Indeed, recent studies in [2][3] show that F_{b^3} is a superior metric to the classical F_β measures to compute clustering accuracy.

The obtained results show interesting situations. In all cases, Dual *C*-means outperforms state-of-the-art algorithms in terms of clustering accuracy. In particular, SCP, DICE and LogLike show improved results and outperform other word-word similarity metrics. It is interesting to notice that PMI and Φ^2, which are known to give less importance to more frequent events show less relevant results. As for the state-of-the-art algorithms, best results are obtained by STC improving over TOPICAL and LDA.

These results only give a small idea of the overall phenomena. In Tables 4, 5 and 6, results for 10 cluster outputs are given for all metrics and all datasets. These new results show interesting properties of evaluation metrics. Although Dual *C*-means shows improvements over all competitors in terms of F_{b^3} or F_1^C (except in one case) for ODP-239, SEMEVAL and WEBSRC401, this situation does not stand for the other metrics, ARI or F_1^N. For ODP-239,

[15] http://gate.ac.uk/ [Last access: 27/01/2014].

[16] Note that for our experiments, the first seed Web snippet is selected as the one, which is most similar to all other ones in S.

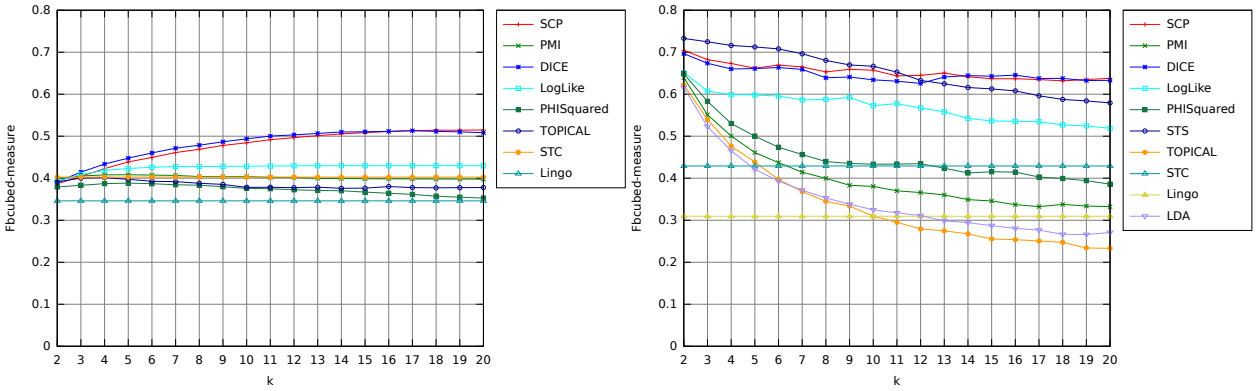

Figure 3: Impact of K for F_{b^3} against ODP-239 (left) and WEBSRC401 (right) datasets.

the best results are obtained by LDA in terms of ARI and LINGO in terms of F_1^N. For SEMEVAL, the best performances are provided by STC in terms of ARI and LINGO in terms of F_1^N. Deep analysis shows that ARI embodies an interesting property for the SRC task as it is well-known that the sizes of the clusters are not distributed equally on the Web. Indeed, ARI tends to favor solutions, which show similar partitioning shapes to the gold standard. As a consequence, a good SRC system should be performant both in terms of ARI and F_{b^3}. On the other hand, F_1^N shows inconsistent results when compared to all other metrics. In particular, it tends to give high results when the other metrics decrease.

Although different results are obtained for SEMEVAL and ODP-239, steady results are obtained for WEBSRC401 by the Dual C-Means configured with the S^TS word-word similarity metric. Indeed, it clearly outperforms all other algorithms in terms of F_{b^3}, F_1^C and ARI. At this stage of our experiements, we can conclude that this configuration provides the best performance both in terms of clustering accuracy and partitioning shape.

		F_1^N	*ARI*	F_{b^3}	F_1^C
ODP-239	LDA	0.5978	**0.2571**	**0.4370**	**0.3900**
	LINGO	**0.6636**	0.0920	0.3461	0.2029
	STC	0.5499	0.1597	0.4027	0.3238
	TOPICAL	0.5760	0.1505	0.3799	0.2839
SEMEVAL	LDA	0.7159	0.1313	0.3966	0.2840
	LINGO	**0.7742**	0.0783	0.3662	0.2072
	STC	0.7223	**0.1704**	**0.4632**	**0.3682**
	TOPICAL	0.6791	0.0621	0.3998	0.2723
WEBSRC	LDA	0.7020	**0.0268**	0.3214	0.2613
	LINGO	**0.7123**	0.0247	0.3095	0.2502
	STC	0.6779	0.0220	**0.4293**	**0.3905**
	TOPICAL	0.6932	0.0203	0.3083	0.2522

Table 4: Results of state-of-the-art algorithms for the ODP-239, SEMEVAL, WEBSRC401. K fixed to 10 Clusters for LDA and TOPICAL.

The second set of our experiments aims to analyse the behavior of Dual C-Means when external resources are included. In this case, we use the set of query logs provided by the NTCIR-10 Intent-2 task [32] and propose to drive the clustering process by this external information. As such, a cluster centroid is represented by its most representative query log. Results are presented in Table 6 where $V_1 \neq V_2$ for WEBSRC401. Let us notice that this is the only dataset for which experiments with query logs can be performed and easily reproduced.

		F_1^N	*ARI*	F_{b^3}	F_1^C
SEMEVAL	SCP	0.6114	0.0435	0.5632	0.4856
	PMI	0.6634	**0.1072**	0.4198	0.3297
	DICE	0.6245	0.0545	**0.5763**	0.4914
	LOGLIKE	0.5753	0.0209	0.5416	**0.4934**
	Φ^2	**0.6797**	0.1055	0.3972	0.2932
	S^TS	0.6225	0.0319	0.5722	0.4808
	LSA	0.6219	0.0240	0.5645	0.4684
ODP239	SCP	0.4961	0.0865	0.4845	0.3785
	PMI	**0.5671**	**0.1741**	0.4041	0.3231
	DICE	0.5181	0.1213	0.4939	**0.3885**
	LOGLIKE	0.5078	0.1388	0.4285	0.3650
	Φ^2	0.5479	0.1618	0.3759	0.3059
	S^TS	0.5294	0.1304	**0.4852**	0.3822
	LSA	0.5482	0.1490	0.4712	0.3731

Table 5: Results of the Dual C-Means algorithm for ODP-239 and SEMEVAL. K fixed to 10 Clusters. Let us notice that for all experiments, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets.

		F_1^N	*ARI*	F_{b^3}	F_1^C
$V_1 = V_2$ (Text)	SCP	0.6698	0.0317	0.6597	0.6217
	PMI	0.6788	0.0280	0.3981	0.3514
	DICE	0.6718 †	0.0341	0.6575	0.6202
	LOGLIKE	0.6566	0.0242	0.5499	0.5131
	Φ^2	**0.6841**	0.0213	0.4299	0.3836
	S^TS	0.6713	**0.0343**	**0.6666** †	**0.6260** †
	LSA	0.6706	0.0170	0.6327 †	0.5884 †
$V_1 \neq V_2$ (QL)	SCP	0.6580	0.0418	**0.6572**	**0.6239**
	PMI	**0.6866**	0.0366	0.3806	0.3338
	DICE	0.6593	0.0320	0.6343	0.6023
	LOGLIKE	0.6636	0.0219	0.5728	0.5394
	Φ^2	0.6783	0.0267	0.4333	0.3926
	S^TS	0.6645	**0.0470**	0.6160	0.5847
	LSA	0.6719	0.0403 †	0.5577	0.5264

Table 6: Results of the Dual C-Means algorithm for WEBSRC401. K fixed to 10 Clusters. Let us notice that for all experiments where $V_1 = V_2$, the number of p words composing the centroids was set to 2 and the vocabulary is the set of words appearing in the retrieved Web snippets. Note that † means paired student's t-test statistical relevance for $p - value < 0.05$ between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$.

Not surprinsingly, the introduction of external information decreases clustering accuracy. But, this is true only for a glimpse when comparing $S^T S$ for $V_1 = V_2$ and SCP for $V_1 \neq V_2$ (statistical relevance is not true in this case). However, the difference in terms of ARI is higher in favor of the dual representation space, although not with statistical relevance. In this case, we can conclude that while clustering accuracy slightly drops, partitioning shape seems to be put in advance by the query log driven approach. The other benefit of this new dual approach may be embodied by the expressiveness of the query logs as meaningful labels. This is the objective of the next section.

6. LABELING EVALUATION

As mentioned in [9], the labeling process plays an important role in the success of SRC systems. As a consequence, a clear objective evaluation is needed. However, this has not yet been the case. Indeed, [18][16] proposed user studies, which are difficult to replicate. In order to solve reproducibility problems, [10][35] proposed to evaluate the *k*SSL metric but their datasets are defined in two different ways and they are not publicly available. So, in order to propose a conclusive evaluation of the labeling process, we propose to use a new gold standard dataset provided by the Subtopic Mining subtask of the NTCIR-10 Intent-2 [32] and apply recent evaluation metrics proposed by [34]: *I-rec*@10, *D-nDCG*@10 and *D#-nDCG*@10.

These metrics aim to measure Precision and Recall of the users' intents. Within our context, we can use the labels provided by the SRC algorithms as the users' intents candidates. If so, we can directly apply the given metrics. In particular, *I-rec* measures the number of intents discovered by the algorithm over the total different intents of the query. This metric can simply be viewed as an intent Recall. Then, *D-nDCG* is obtained by sorting all relevant intents by the global gain, which is defined as the sum of all the individual intent gains. Finally, the *D#-nDCG* metric is the linear combination of *I-rec* and *D-nDCG*, using γ and $1-\gamma$ factors. Note that defining the probabilities of each intent as well as the relevant intents can be a hard task. However, as our experiments are realized over WEBSRC401 based on ClueWeb09, these values are known and publicly available [32]. In particular, the NTCIREVAL toolkit[17] was used for the calculation of these metrics. Let us notice that for the specific task of SRC, we propose to use *I-rec*@10, *D-nDCG*@10 and *D#-nDCG*@10 as for most queries the number of intents is limited. These metrics are defined in the Equations 7, 8 and 9.

$$I\text{-}rec@N = \frac{|I'|}{|I|} \tag{7}$$

where I is the set of known intents for a query q and I' is the set of intents covered by the returned labels at level N.

$$D\text{-}nDCG@N = \frac{\sum_{r=1}^{N}\sum_i Pr(i|q)g_i(r)/log(r+1)}{\sum_{r=1}^{N}\sum_i Pr^*(i|q)g_i^*(r)/log(r+1)} \tag{8}$$

where $Pr(i|q)$ (resp. $Pr^*(i|q)$) denotes the intent probability obtained for the discovered labels (resp. for the reference labels) and $g_i(r)$ (resp. $g_i^*(r)$) is the gain value of the label at rank r with respect to i for the output of the labeling (resp. for the reference labeling).

[17]http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html [Last access: 27/01/2014].

$$D\#\text{-}nDCG@N = \gamma I\text{-}rec@N + (1-\gamma)D\text{-}nDCG@N \tag{9}$$

where γ was set to 0.5 following the framework evaluation proposed in the Subtopic Mining subtask of the NTCIR-10 Intent-2.

The results provided by [33] for different query completions ($Bing_C$, $Google_C$ and $Yahoo_C$), query suggestions ($Bing_S$) services and a simple merging strategy (Merge) are reported in Table 7 as well as the results of our approach. In particular, we show the results when clustering is query log driven ($V_1 \neq V_2$) and when labeling is performed *a posteriori* ($V_1 = V_2$). By *a posteriori*, we mean that clustering is first performed on the exclusive text representation. Then, as a usual second step, the label is computed by any heuristic. In our experiments, the query log that best represents each text-based cluster is computed using one iteration of the update function defined in section 3, which allows direct comparison results.

		$I-rec@10$	$nDCG@10$	$D\#-nDCG@10$
$V_1 = V_2$	SCP	0.2804	0.3195	0.2959
	PMI	0.3136	0.3444	0.3250
	DICE	0.2952	0.3242	0.3093
	LOGLIKE	0.2269	0.2885	0.2550
	Φ^2	0.3390	0.3642	0.3523
	$S^T S$	0.2837	0.3063	0.2935
	LSA	0.3238	0.3694	0.3456
$V_1 \neq V_2$	SCP	0.3669 †	0.3932 †	0.3793 †
	PMI	0.4136 †	0.4257 †	0.4203 †
	DICE	0.3761 †	0.3884 †	0.3814 †
	LOGLIKE	0.3937 †	0.4146 †	0.4046 †
	Φ^2	**0.4249** †	0.4221 †	**0.4225** †
	$S^T S$	0.4033 †	**0.4273** †	0.4119 †
	LSA	0.3946†	0.4197 †	0.4050 †
Baselines	BingS	0.3068	0.2787	0.2928
	BingC	0.3231	0.3268	0.3250
	GoogleC	0.3735	0.3841	0.3788
	YahooC	0.3829	0.3815	0.3822
	Merge	0.3365	0.3181	0.3273

Table 7: Evaluation results of the labeling process with query logs over the NTCIR-10 Intent-2 dataset. Note that † means paired student's t-test statistical relevance for $p-value < 0.05$ between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$.

The results of the query driven Dual *C*-Means outperform all baselines and *a posteriori* labeling. Moreover, all the differences between a given metric in $V_1 = V_2$ and its counterpart in $V_1 \neq V_2$ are statistically relevant. These results also show interesting behaviors. Indeed, while PMI and Φ^2 collocation metrics previously showed worst clustering accuracy results compared to other configurations, they show improved results in terms of labeling. The fact that these metrics tend to favour less frequent associations is an interesting characteristic for labeling purposes and a conclusive remark. Moreover, the $S^T S$ word-word similarity measure shows high $nDCG@10$ value and competitive overall $D\#-nDCG@10$. These results clearly point at this last configuration as the best compromise for clustering accuracy, labeling quality and partitioning shape.

7. CONCLUSIONS AND PERSPECTIVES

In this paper, we proposed a new algorithm called Dual *C*-Means, which can be seen as an extension of the classical *K*-Means for dual representation spaces. Its originality relies in the fact that the clustering process can be driven by external resources by defining two distinct representation spaces. In particular, we proposed

that query logs are used as external information to guide clustering and afford meaningful labels to users in their search for information. We also built a new publicly available dataset called WEB-SRC401 based on ClueWeb09, which affords a more realistic situation for Web SRC. A complete and reproducible evaluation was performed over different gold standard datasets (ODP-239 and SEMEVAL) based on different publicly available evaluation tools. In particular, a great deal of evaluation metrics have been applied over diffferent configurations of the Dual *C*-Means integrating distinct word-word similarity measures. Results showed that our approach steadily outperforms all existing state-of-the-art SRC algorithms in terms of clustering accuracy (F_{b^3}) but is less competitive in terms of ARI. This situation is handled by the introduction of query logs, which allows high labeling quality with outperforming values of $I-rec@10$, $D-nDCG@10$ and $D\#-nDCG@10$ and adequate partitioning shape with high values of ARI.

The final findings that show that collocation metrics sensitive to high frequency events tend to produce high quality clusters and low frequency sensitive ones give rise to quality labels, is an interesting issue. Indeed, like the dual representation space, it suggests a multiobjective implementation of the dynamic reallocation algorithm to the problem of SRC. Moreover, the next steps that are being carried out are the introduction of different resources to drive the clustering process, the definition of new *P* transition matices taking into account recent developments in word-word similarity and the definition of powerful instantiation functions provided by the introduced general model.

8. REFERENCES

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 97–106, 2009.

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

[3] E. Amigó, J. Gonzalo, and F. Verdejo. 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, 2013.

[4] D. Arthur and S. Vassilvitskii. K-means++: the advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

[5] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. Scalable k-means++. *Proceedings of the Very Large Data Base Endowment (PVLDB)*, 5(7):622–633, 2012.

[6] A. Bernardini, C. Carpineto, and M. D'Amico. Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 206–213, 2009.

[7] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] T. Brants and A. F. Web 1t 5-gram, 2006.

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

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

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

[12] A. Di Marco and R. Navigli. Clustering web search results with maximum spanning trees. In *Proceedings of the 12th International Conference on Artificial Intelligence Around Man and Beyond (AI*AI)*, pages 201–212, 2011.

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

[14] L. Dice. Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302, 1945.

[15] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

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

[17] W. Gale and K. Church. Concordances for parallel texts. In *Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, pages 40–62, 1991.

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

[19] W. Kong and J. Allan. Extracting query facets from search results. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 93–102, 2013.

[20] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, pages 211–240, 1997.

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

[22] Y. Maarek, R. Fagin, I. Ben-Shaul, and D. Pelleg. Ephemeral document clustering for web applications. Technical report, IBM, 2000.

[23] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[24] D. Milne and I. Witten. An open-source toolkit for mining wikipedia. *Journal of Artificial Intelligence*, 194:222–239, 2013.

[25] J. Moreno and G. Dias. Using text-based web image search results clustering to minimize mobile devices wasted space-interface. In *Proceedings of 35th European Conference on Information Retrieval (ECIR)*, pages 532–544, 2013.

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

[27] R. Navigli and G. Crisafulli. 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, 2010.

[28] R. Navigli and D. Vannella. Semeval-2013 task 11: Word sense induction & disambiguation within an end-user

application. In *Proceedings of the International Workshop on Semantic Evaluation (SEMEVAL)*, pages 1–9, 2013.

[29] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 15th Neural Information Processing Systems Conference (NIPS)*, pages 849–856, 2001.

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

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

[32] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. Kato, R. Song, and M. Iwata. Summary of the ntcir-10 intent-2 task: Subtopic mining and search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 761–764, 2013.

[33] T. Sakai, Z. Dou, T. Yamamoto, M. Lui, Y. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In *Proccedings of the Research Infrastructure for Comparative Evaluation of Information Retrieval and Access Technologies (NTCIR)*, 2013.

[34] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th international ACM conference on Research and development in Information Retrieval (SIGIR)*, pages 1043–1052, 2011.

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

[36] J. Silva, G. Dias, S. Guilloré, and J. Lopes. 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, 1999.

[37] N. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1073–1080, 2009.

[38] D. Vitale, P. Ferragina, and U. Scaiella. Classification of short texts by deploying topical annotations. In *34th European Conference on Advances in Information Retrieval (ECIR)*, pages 376–387, 2012.

[39] X. Wang, W. Gu, D. Ziebelin, and H. Hamilton. An ontology-based framework for geospatial clustering. *International Journal of Geogaphical Information Science*, 24(11):1601–1630, 2010.

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

[41] D. Zhang and Y. Dong. Semantic, hierarchical, online clustering of web search results. In *Proceedings of the 6th Asia Pacific Web Conference (APWEB)*, pages 69–78, 2004.