UNSUPERVISED ORGANISATION OF SCIENTIFIC DOCUMENTS

André Lourenço[1,2], Liliana Medina[3], Ana Fred[2] and Joaquim Filipe[3,4]
[1]*Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal*
[2]*Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal*
[3]*Institute for Systems and Technologies of Information, Control and Communication, Lisbon, Portugal*
[4]*School of Technology of Setúbal, Polytechnic Institute of Setúbal, Setúbal, Portugal*

Keywords: Unsupervised learning, Clustering, Clustering combination, Clustering ensembles, Text mining, Feature selection, Concept induction, Metaterm.

Abstract: Unsupervised organisation of documents, and in particular research papers, into meaningful groups is a difficult problem. Using the typical vector-space-model representation (Bag-of-words paradigm), difficulties arise due to its intrinsic high dimensionality, high redundancy of features, and the lack of semantic information. In this work we propose a document representation relying on a statistical feature reduction step, and an enrichment phase based on the introduction of higher abstraction terms, designated as metaterms, derived from text, using as prior knowledge papers topics and keywords. The proposed representation, combined with a clustering ensemble approach, leads to a novel document organization strategy. We evaluate the proposed approach taking as application domain conference papers, topic information being extracted from conference topics or areas. Performance evaluation on data sets from NIPS and INSTICC conferences show that the proposed approach leads to interesting and encouraging results.

1 INTRODUCTION

The increase in the volume of scientific literature and its dissemination using the Web is leading to an information overload. Scientific literature comprises different kinds of publications, as scientific articles published in journals, book chapters, papers in conferences, technical reports, etc. This kind of literature has a standardized structure (title, abstract, introduction, methods, results, conclusions), and typical writing style, resulting in very similar documents. One of the key problems is that it is poorly categorized, being difficult to retrieve all the main articles of a specific topic.

Autonomous citation indexing (ACI) (Lawrence et al., 1999) has been proposed to help the organization of scientific literature by automating the construction of citation indices. A citation index catalogues the citations that an article makes, linking the articles with the cited works. These mechanisms help scientists to find work that cites their own work or is relevant to their research, but does not solve the problem of documents organization.

Machine learning methods have been used proposing several approaches for the problem. Document clustering provides a possible solution, grouping articles into categories, based on different information extracted from them, using only the textual content of the article (Janssens et al., 2006)(Aljaber et al., 2010), and/or the citation graph analysis (Ahlgren and Jarneving, 2008; Boyack and Klavans, 2010).

The representation of the textual content of documents using the standard bag-of-words model is only effective for grouping related documents when sharing a large proportion of lexically equivalent terms. This standard approach ignores synonymies and other relations between words, which reduces the effectiveness of this document representation scheme.

In this work we propose a methodology to cluster scientific literature based on its textual content and on a priori categorization of each document on broad classes according to a classification system provided to the authors. We propose an extension of the typical bag-of-words representation introducing metaterms, concepts derived from the text that try to go beyond the syntax, forming a conceptualization that connects related terms.

We follow a recent and promising trend in unsupervised learning, namelly clustering combination techniques (Fred, 2001; Fred and Jain, 2005; Strehl

Lourenço A., Medina L., Fred A. and Filipe J..
UNSUPERVISED ORGANISATION OF SCIENTIFIC DOCUMENTS.
DOI: 10.5220/0003722905490560
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (SSTM-2011), pages 549-560
ISBN: 978-989-8425-79-9

and Ghosh, 2002; Hanan and Mohamed, 2008), which typically outperform the results of single clustering algorithms, achieving better and more robust partitioning of the data, combining the information provided by a *clustering ensemble* (CE). Moreover this class of algorithms further enables the combination of several sources of information (e.g. citations) for the clustering of scientific documents.

The remainder of the paper is organized as follows. Section 2 describes related work. Section 3 presents the proposed methodology, dividing the solution in the following phases: stop-word removal (in section 3.1); meta-terms creation (in section 3.2); and clustering strategy (in section 3.3). Section 4 presents the data sets used for evaluation and section 5 the main results. Finally, in section 6 we draw the main conclusions and future work.

2 RELATED WORK

Document categorization can be divided in two stages: a) transforming the documents into a suitable and useful data representation; b) organizing the data into meaningful groups.

Consider a set of D documents $X = \{d_1, \ldots, d_D\}$ to be clustered. Commonly, document information is represented using a vector space model or bag-of-words (Manning et al., 2008) where each document, d_i, is represented by document tokens aggregated in a feature vector with F dimensions, $d_i = \{w_1, \ldots, w_F\}$; w_j, represents the relative importance of each token in the document. Typically w_j is computed using the Term Frequency-Inverse Document Frequency weighting (TF-IDF) (Sebastiani, 2005). The most simple form of a token is a word, but compound terms can also be used such as bigrams, trigrams and noun phrases.

Recent work (Hotho et al., 2003; Sedding and Kazakov, 2004; Reforgiato Recupero, 2007), considers not only syntactic information, obtained from the terms present in a document, but also semantic relationships between terms. These approaches are mostly based on WordNet (Fellbaum, 1998), which is a lexical database that groups English words into sets of synonyms, called synsets. Most synsets are connected to other synsets via several semantic relations, such as: synonymy, hypernymy, meronymy and holonymy.

In (Hotho et al., 2003), the standard representation of a document is enriched with concepts derived from WordNet, using several strategies: (a) replacing terms by concepts; (b) using concepts only; (c) or extending the term vector with WordNet concepts. These strategies lead to a new document representation, d_i', which includes a concept subspace, $d_i' = \{w_1, \ldots, w_F, c_1, \ldots c_C\}$, where c_i represents a concept.

These concepts correspond to hypernyms of extracted terms, that is to more generic words than the extracted term. In (Hotho et al., 2003), the analysis relied on single-term analysis, while in (Zheng et al., 2009b) a more complete phrase-based analysis was performed.

Concerning the clustering algorithm several approaches are followed in the literature. In (Hotho et al., 2003; Sedding and Kazakov, 2004; Reforgiato Recupero, 2007) a variant of the K-means, the Bi-Section-K-means is used, stating that this method frequently outperforms the standard K-means. In (Boyack et al., 2011) a more complex partitioning of the document collection is proposed. They start by obtaining a graph of related documents, based on several document similarity measures; after employing a pruning step over the obtained graph, an average-link algorithm is used to assign each document to a cluster based on proximity of remaining edges. This process is run 10 separate times with different starting points, and the results are re-clustered, or combined, using only those document pairs that are clustered together in at least 4 of 10 preliminary solutions. The method was evaluated on more than two million biomedical publications.

3 PROPOSED APPROACH

The motivation behind the proposed methodology consists on the unsupervised organization of scientific papers into meaningful subsets, accomplished through the derivation of a conceptualization relating terms that describe the categories used by authors to classify their articles using a scientific classification system, such as the ACM Computing Classification System (ACM, 1998). This conceptualization is created using a bottom-up approach, based on documents' textual content. This information can be interpreted as *a priori* knowledge, and allows a more accurate representation of documents content, improving the typical bag-of-words representation.

Our approach differs from previous work, in two perspectives: (a) the analysis of text relies on compound terms; (b) we add hypernyms to the base representation, which are not imposed by Wordnet, but are instead extracted from the data textual content.

Figure 1 presents a general overview of the proposed methodology.

We start by the extraction of single words, w_i, and compound terms, ct_j (bigrams, trigrams and/or noun

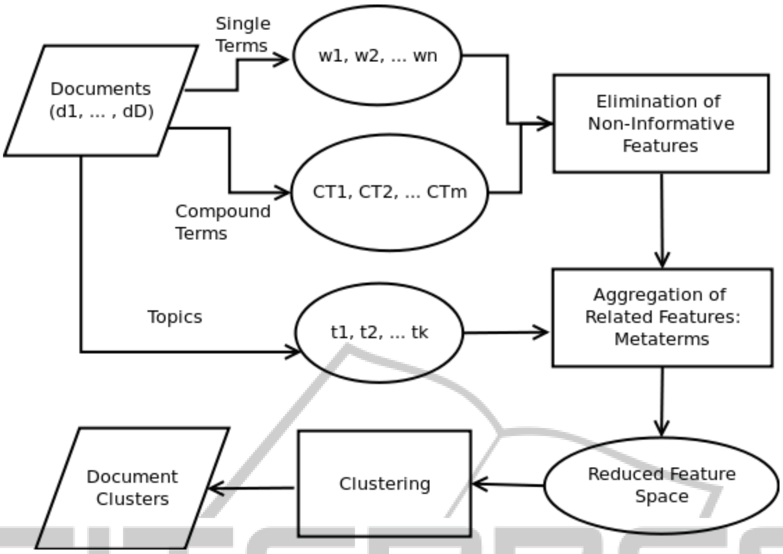


Figure 1: Proposed Methodology: (1) Document Representation; (2) Elimination of Non-Informative Features; (3) Aggregation of Related Features; (4) Clustering.

phrases) from the text; this takes place after removing all the punctuation from the text, as well as mathematical formulas. These features are extracted from the documents using a natural language processing (NLP) tool specifically built for this purpose. Each document is represented by a feature space, composed by, $d_i'' = \{w_1, \ldots, w_M, ct_1, \ldots ct_M\}$.

This representation has the following challenges:

- The dimensionality $(w+m)$ is very high.
- Existence of redundant or irrelevant features.
- This representation does not take into account the possible, semantic, relationships between features.

In order to reduce the feature space, we propose two independent steps: (1) Removal of non-informative words in the context of the document collection; (2) Aggregation of related features (both words and compound terms) into a more general and more meaningful feature, which we refer to as metaterm.

Our hypothesis is that the feature space reduction by aggregating terms into metaterms will provide a better, more accurate text representation.

Using this alternative representation, we finally organize the documents using a clustering combination algorithm.

3.1 Context-dependent Stop-Word Removal

Stop word removal is one of the most common approaches to non-informative feature removal. Stop-words are terms that appear too frequently in documents and thus provide low discriminative power (Van Rijsbergen, 1979). In this work, we go one step further, assuming that this problem is context-dependent. We address this problem applying a statistical criterion to the document collection in analysis, trying to remove redundant or irrelevant features of the type word: the percentage of documents where the word occurs.

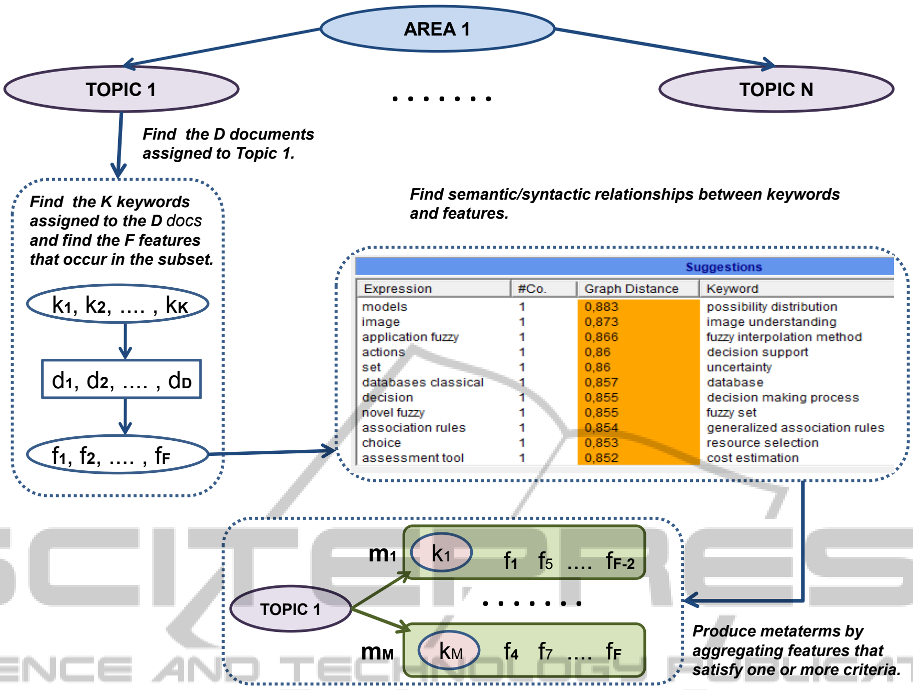
We consider that if a word occurs in a very high percentage of documents, then it must not be very meaningful in the context of the data set. The same hypothesis is applied for the case when the term occurs in a very small percentage of documents.

After applying this step over a data set, we expect to obtain a reduced feature space comprised of the most meaningful words and compound terms present in the text.

3.2 Aggregation of Related Features: MetaTerm Creation

In (Zheng et al., 2009a), concept induction is performed using Latent Semantic Analysis (LSA) techniques, based on single terms and common phrases extracted from the documents. Relations between concepts are constructed using WordNet, extracting hypernyms for each detected concept. The drawback of this approach is the dependance on WordNet, which is a general and context-independent resource.

We define a metaterm as the entity representing a


AREA 1
TOPIC 1
....
TOPIC N
Find the D documents assigned to Topic 1.
Find the K keywords assigned to the D docs and find the F features that occur in the subset.
k1, k2, ... , kK
d1, d2, ... , dD
f1, f2, ... , fF
Find semantic/syntactic relationships between keywords and features.
m1
k1
f1 f5 .... fF-2
TOPIC 1
....
mM
kM
f4 f7 .... fF
Produce metaterms by aggregating features that satisfy one or more criteria.

Suggestions

Expression	#Co.	Graph Distance	Keyword
models	1	0,883	possibility distribution
image	1	0,873	image understanding
application fuzzy	1	0,866	fuzzy interpolation method
actions	1	0,86	decision support
set	1	0,86	uncertainty
databases classical	1	0,857	database
decision	1	0,855	decision making process
novel fuzzy	1	0,855	fuzzy set
association rules	1	0,854	generalized association rules
choice	1	0,853	resource selection
assessment tool	1	0,852	cost estimation

Figure 2: Aggregation of related features.

subset of words and/or compound terms which are either synonyms or have some kind of semantic relation (with variable degrees of ‘closeness’). These concepts have a direct analogy to Wordnet hypernyms, since they represent also higher-level expressions, but are obtained using different sources.

Our approach uses as external source of *a priori* information the conference topics, assigned by authors when submitting their work to conferences. These topics are considered the roots of the metaterms to be created. For each topic we analyse the keywords provided for each document.

We propose as criterion for the extraction of metaterms based on the keywords an adaptation of the Lesk Algorithm (Dao and Simpson, 2005) (Banerjee and Pedersen, 2003). The Lesk Algorithm (Lesk, 1986) disambiguates words in short phrases, comparing the dictionary definition or gloss of a given word to the glosses of every other word in the phrase. A word is assigned the sense whose gloss shares the largest number of words with the glosses of the other words. The used adaption consists of using WordNet as dictionary, with senses arranged in a hierarchical order. This criterion was chosen based on the assumption that words in a given neighborhood will tend to share a common topic.

Figure 2 presents these procedures, showing for a given topic an example of terms/compound terms related with keywords that will produce a metaterm. The minimum number of co-occurrences/ Lesk Distance is specified by the user.

An example of three metaterms obtained for the topic ”‘Datamining”’ may be observed in Figure 3. Each metaterm was based on a particular keyword, assigned to documents that belong to the chosen topic (circled purple).

The reduced feature space represents each document as $d_i^{new} = \{m_1, \ldots, m_R, w_1, \ldots, w_T, ct_1, \ldots ct_L\}$, where m_i represents a metaterm, and w_j and ct_j single terms and compound terms that were not aggregated.

We used the TF-IDF weighting scheme. Let’s denote $tf_{t_i,j}$ as the frequency of the feature t_i on the document d_j, when t_i is either a word or a compound term; and idf_{t_i} as the the inverse frequency of t_i.

The TF-IDF weight for a metaterm, m_i is a combination of the frequency values of its components. Let T be the number of components of m_i. The number of occurrences of this metaterm in a document d_j, $\#m_{i,j}$ is given by $\#m_{i,j} = \sum_{t \in T} \#_{t,j}$ where $\#_{t,j}$ is the number of occurrences of each term $t \in T$ within d_j. Let M be the subset of metaterms that occur in d_j: thus, the frequency of m_i in the document is $tf_{m_i,j} = \frac{\#m_{i,j}}{\sum_{q \in M} \#m_{q,j}}$. The inverse frequency of m_i is $idf_{m_i} = log \frac{D}{\{d : m_i \in D\}}$, and finally

$$tfidf_{m_i,j} = tf_{m_i,j} \times idf_{m_i} \qquad (1)$$

The feature space may be further (or alternatively) reduced by applying techniques such as Latent Se-

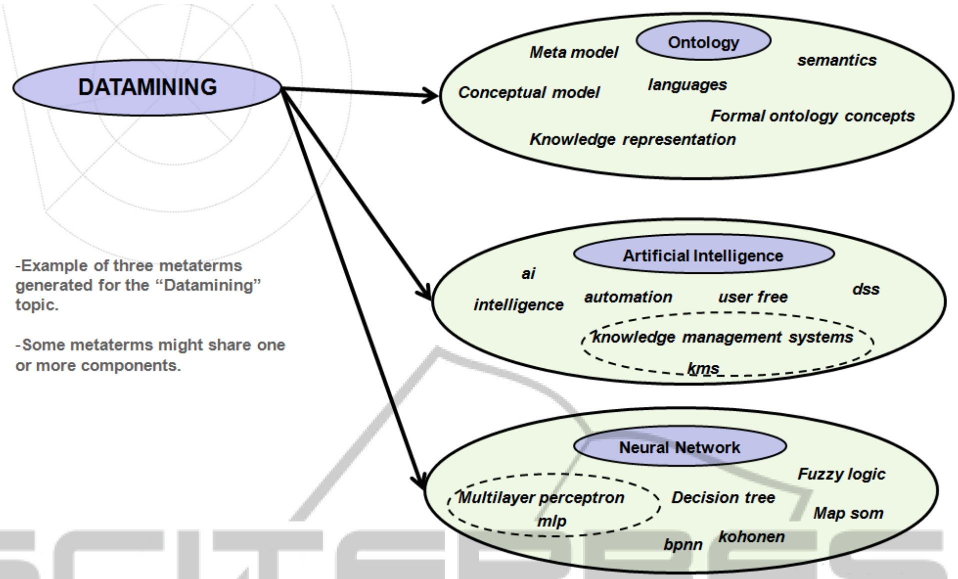


Figure 3: Three metaterms obtained for the topic "'Datamining'", based on the keywords "Ontology", "Artificial intelligence" and "Neural network".

mantic Indexing (LSI), which reduces the number of features by generating a new feature space that would capture, ideally, the true (*semantic*) relationships between documents(Sevillano et al., 2009). It uses Single Value Decomposition (SVD) to decompose the TF-IDF matrix into a subset set of k orthogonal vectors:

$$M = U\Sigma V^T, \tag{2}$$

where M is a term-by-document matrix, U and V^T are the left and right singular vectors matrices, and σ is the singular values matrix. Dimensionality reduction is accomplished by retaining the first z columns of matrix V (Sevillano et al., 2009). Therefore, the dimensionality of the reduced feature space, z, must be carefully chosen.

3.3 Clustering

A clustering algorithm organizes a set of objects, in this case documents, into k clusters by generating a partition of the data into k groups, $P = \{C_1,\dots,C_k\}$. The assignment of these patterns into different clusters is based on a given similarity or dissimilarity metric such that the similarity between patterns of the same cluster is greater than the similarity between patters belonging to different clusters.

Different clustering algorithms lead in general to different organization of patterns. A recent approach consists on the production of a more robust clustering results by combining the results of different partitions, called the clustering ensembles (CE). A CE is a set of N different partitions of X, $\mathbb{P} = \{P^1,\dots,P^N\}$, where each partition $P^i = \{C^i_1,\dots,C^i_{k_i}\}$, has k_i clusters. This partitions can be generated by the choice of clustering algorithms or algorithmic parameters, or using different feature representations, as described in (Fred and Jain, 2005).

Evidence Accumulation (EAC) is one of the clustering ensemble methods that enables the combination of several partitionings of the data set. The underlying assumption is that patterns belonging to a natural cluster are very likely to be assigned in the same cluster in different partitions. Taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the N data partitions of n patterns are mapped into a $n \times n$ co-association matrix:

$$C(i,j) = \frac{n_{ij}}{N} \tag{3}$$

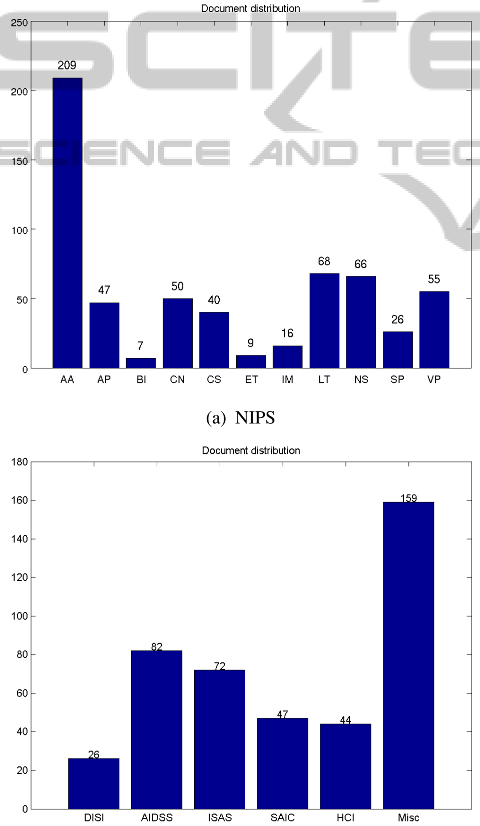
where n_{ij} is the number of times the pattern pair (i,j) is assigned to the same cluster among the N partitions. This matrix corresponds to an estimate of the probability of pairs of objects belonging to the same group, as assessed by the N partitions of the ensemble.

A consensus partition can be extracted applying a clustering algorithm, which typically induces a hard partition, to the co-association matrix (Fred and Jain, 2005). The decision on the number of clusters of the consensus partition, might be based on specific criteria, such as the cluster lifetime criterion (Fred and Jain, 2005), or based on ground truth information.

In this work, as source for the clustering algorithms we used the different representations of documents described before. The construction of the clustering ensemble (CE) is generated using the k-means algorithm, taking as distance measure one minus the cosine of the angle between the vectors representing the documents. Each partition of the ensemble is produced with a varying number of clusters. The minimum and the maximum number of clusters were determined as a function of the number of samples, n_s,as given by the expression (Lourenço et al., 2010):

$\{K_{min}, K_{max}\} = \{\lceil n_s/A \rceil, \lceil n_s/B \rceil\}$, with $A = 50$ and $B = 20$.

For the extraction of the consensus partition we used several hierarchical agglomerative methods (Single Link - SL, Complete Link - CL, Average Link - AL, Ward's Link - WL), and Metis algorithm (Karypis et al., 1998). The number of extracted clusters is equal to the number of topics of each conference.



(a) NIPS

(b) ICEIS

Figure 4: Documents distribution: 4(a) document-per-topic distribution of the NIPS papers, and 4(b) document-per-area distribution of the ICEIS papers.

4 EXPERIMENTAL SETUP

We applied the previously detailed methodology to two data sets, referred to as NIPS and ICEIS.

The NIPS data set, built by (Globerson et al., 2007), consists of 2484 documents from 17 NIPS conferences held between 1987 and 2003. There are 14036 distinct words occurring in this data set. Given that the topic distribution for 593 documents from years 2001 to 2003 is available, we applied our methodology to this subset of papers. The original feature space for this subset is comprised of 6881 words.

The ICEIS[1] data set consists of 430 documents from one conference organized by INSTICC[2] in 2009. The ICEIS event is subdivided into 5 conference areas, and each area is further subdivided into topics. Documents were grouped according to the area / topic, in a total of 75 topics. This data set contains 20460 distinct features (words, bigrams, trigrams, etc.).

Each NIPS topic has an associated set of expressions [3], assigned by the conferences organizers. Given that the papers submitted to NIPS conferences do not have keywords assigned by the authors, we use these expressions as keywords and search for documents where they occur in order to produce suggestions for metaterms.

The ICEIS keywords-per-document assignment is available, which allows us to combine this information with the documents-per-topic information and build specific metaterms.

Figure 4 depicts the topic distribution for both data sets is. The topics for the NIPS data are: Algorithms & Architectures (AA); Applications (AP); Brain Imaging (BI) ; Control and Reinforcement Learning (CN); Cognitive Science (CS); Emerging Technologies (ET); Implementation (IM); Learning Theory (LT); Neuroscience (NS); Speech and Signal Processing (SP); Vision Processing (VP). The AA category is the largest with 209 documents, representing almost half of the documents of the collection. For the ICEIS data set the topics are: Databases and Information Systems Integration (DISI); Artificial Intelligence and Decision Support Systems (AIDSS); Information Systems Analysis and Specification (ISAS); Software Agents and Internet Computing (SAIC); Human-Computer Interaction (HCI); Miscelaneous topic, representing documents with more than one topic (Misc). This last category is the largest with 159 documents.

These categorizations are not very clear and many times very fuzzy. Due to this reason, the methodology followed for the evaluation of the results is not based

[1] International Conference on Enterprise Information Systems: http://www.iceis.org/

[2] http://insticc.org/

[3] http://nips.cc

Table 1: Experimental framework: associated parameter values and description.

Experiment	Data Set	Task	Algorithm/Parameter		Description	Value
1,2,3	NIPS and ICEIS	Stopwords Removal	Removal (%)	Max_{SW}	Maximum percentage of documents where a word may occur.	12%
				Min_{SW}	Minimum percentage of documents where a word may occur.	0.5%
1,2,3	NIPS and ICEIS	Clustering Ensemble	K-Means	N_P	number of partitions	200
				k_{min}	minimum number of clusters	$\frac{N_s}{50}$
				k_{max}	maximum number of clusters	$\frac{N_s}{20}$
1,2,3	NIPS and ICEIS	Extraction	SL, CL, AL, WL, metis	k lifetime	Final partition's cluster number	11 and 5
2	NIPS	Feature space reduction	LSI	M	dimensionality of the new feature space	27
				th	threshold	0.4
	ICEIS	Feature space reduction	LSI	M	dimensionality of the new feature space	35
				th	threshold	0.5
3	NIPS	Aggregation	Lesk Algorithm	L_{th}	minimum graph proximity between two features for aggregation	0.9
	ICEIS	Aggregation	Lesk Algorithm	L_{th}	minimum graph proximity between two features for aggregation	0.85

on accuracy calculation (based on this ground truth). We analyse the clusters looking at the features of the documents composing them, to the document distribution, and to the pairwise similarity between the documents on each cluster (obtained from the application of the EAC clustering algorithm).

To understand the impact of the metaterms, we performed, over both data sets, the experiments detailed below. The clustering is always performed using the EAC algorithm over a clustering ensemble of 200 partitions of the baseline representation, obtained with the K-means algorithm.

Experiment 1. Our baseline experiment pertains to the TF-IDF representation matrix obtained after applying the irrelevant feature removal step of the methodology, with no aggregation of terms into metaterms.

Experiment 2. The TF-IDF matrix is transformed by applying LSI over the matrix obtained on Experiment 1.

Experiment 3. Here, the feature aggregation step for metaterm creation is performed in order to reduce the dimensionality of the feature space. We explore two criteria for term aggregation: Lesk algorithm and co-occurrences, obtaining thus two different representations.

The different parameters associated with the experiments are summarized in Table 1. For the Stopwords removal we empirically verify that using as minimum and maximum percentage of documents having a token of 0.5% and 12%, respectively, we conserved words that seemed important for distinguishing documents. For the aggregation of related features step, we empirically chose 0.9 and 0.85 as thresholds for the minimum graph proximity, for the NIPS and ICEIS data set, respectively, trying to guarantee that only strong relations were chosen.

5 RESULTS AND DISCUSSION

Many studies compare cluster solutions based on predefined document sets based on expert opinion. In the present study we do not have such information, or the available information is considered fuzzy. We chose to evaluate the results based on the following: (1) pairwise similarity between documents within clusters, available in the co-association matrices obtained by the EAC clustering algorithm; (2) distribution of documents by topic; (3) Adaptation of within-cluster textual coherence (Boyack and Klavans, 2010), based on Jensen-Shannon divergence; (4) examples of most relevant features of each cluster.

The co-association matrices obtained using EAC, are represented by a color scheme ranging from blue ($\mathcal{C}(i,j)=0$) to red ($\mathcal{C}(i,j)=1$), corresponding to the magnitude of similarity, and the axis represent the documents organized such that documents belonging to the same cluster are displayed contiguously. This information is also represented on the colorbar on top of each figure, where each color represents the obtained clusters. Well formed partitions have a pronounced block-diagonal structure, revealing that the similarity within clusters is very high when compared to documents in different clusters.

Regarding the distribution of documents by topic, we present histograms representing the number of documents assigned to each topic on each of the obtained clusters. Moreover, we present, at the top of the histogram, a bar representing the confidence over each cluster (ranging over the same color scheme as before). This confidence is obtained based on the average similarity of pairwise associations within a cluster.

The textual coherence (Boyack and Klavans, 2010), is computed based on Jensen-Shannon divergence (JSD), which computes the distance between two probability distributions, p and q:

$$JSD(p,q) = 1/2D_{KL}(p,m) + 1/2D_{KL}(q,m) \qquad (4)$$

where $m = (p+q)/2$, and D_{KL} is the Kullback-Leiber divergence

$$D_{KL}(p,m) = \sum(p_i log(p_i/m_i)) \qquad (5)$$

We consider that p and q represent the probabilities of words in two distinct documents. The JSD is calculated for each cluster as the average JSD value over all documents in the cluster, represented as JSD_i. JSD is a divergence measure, meaning that if documents in a cluster are very different from each other its value will be very high; while if documents have very similar words distributions, its value will be low. We obtain the coherence of a partition, as a weighted average over all clusters:

$$Coh = \sum n_i JSD_i / \sum n_i, \qquad (6)$$

where n_i is the number of documents per cluster.
In **Experiment 1** - the baseline experiment, using only the feature removal step, - the NIPS original feature space is reduced to 5660 words, and the ICEIS feature space to 14987 distinct features.

Figure 5 presents the co-association matrices obtained using EAC. When comparing figures 5(a) and 5(b), corresponding to co-association matrices from the NIPS and ICEIS conferences, we see that the first has a more pronounced block-diagonal structure, having clusters apparently more separated. The distribution of documents by topics is represented in figure 6. For the NIPS data set, the majority of the clusters joins several topics (except for clusters 3 and 4); for the ICEIS data set the same happens.

With **Experiment 2**, we obtained a smaller feature space, applying LSI over the TFIDF matrix obtained in the previous experiment. The obtained co-association matrix, for the NIPS data set, has a less obvious block-diagonal structure, with several micro-structures representing small clusters; for the ICEIS data set the cluster structure is more evident.

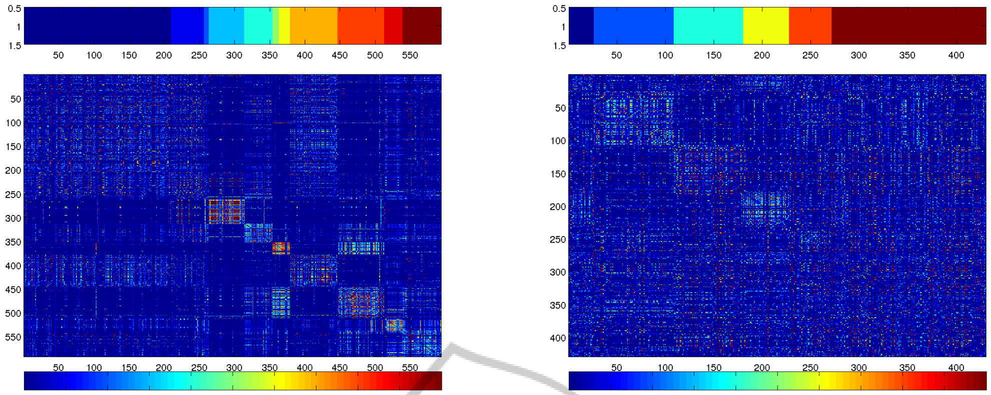
Figure 7 presents the distribution of documents by topics. In the case of NIPS data set, the clusters appear to be more confused, apparently resulting from the grouping of small clusters. In the case of ICEIS, the clusters are better defined than those of experiment 1, with cluster 3 being composed almost entirely by topic 2; nevertheless this partitioning is still very mixed in terms of the distribution of topics per clusters.

Finally, in **Experiment 3**, we summarize the results obtained using metaterms using an adapted version of the Lesk Algorithm. We chose as minimum threshold for aggregating terms normalized distances above 0.9 (ie, normalized distance in the WordNet graph from the keyword that ”generates” the metaterm). From this, 24 metaterms were created for the NIPS data set, and 694 for the ICEIS data set. The reason why ICEIS has a much larger number of metaterms is because information about which keywords were assigned to which papers was made available, unlike the NIPS case. Figure 8 presents the cluster distribution for this experiment. For the NIPS data set, there are several clusters (3,4,5,11) that are composed mainly by one topic. In the case of ICEIS data set, the clusters are better defined than before, with cluster 3,4, and 6, being composed by almost only one topic.

As final note for the graphics of the documents by topics, we observe little correlation between the clusters obtained and the original document-per-topic distribution. This is expected for NIPS data set, due to the fact that some of the conference topics are quite broad (such as ”Algorithms and Applications” or ”Emerging Technologies”); the same is also expected for ICEIS data set given that about 160 documents are assigned to more than one area, which suggests that these are sometimes transversal.

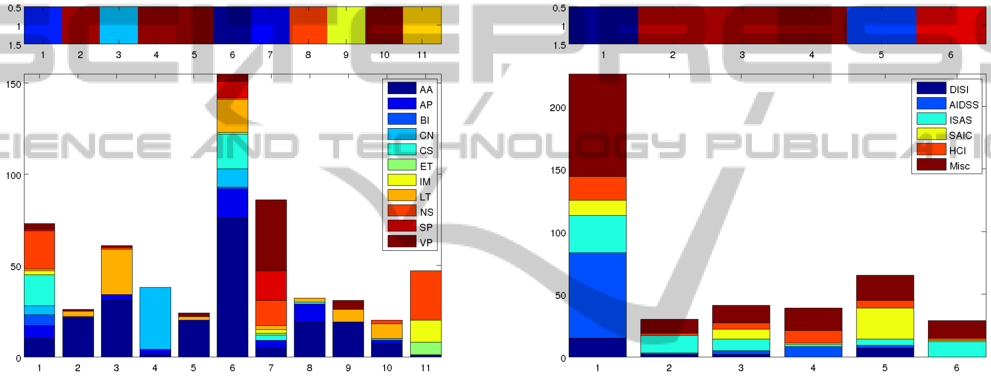
The textual coherences of the different experiments are depicted in figure 9. As one can observe, the lower values of textual coherence are obtained for Experiment 3, using the Lesk criteria (Exp3-Lesk). Regarding other experiments, the presented values follow the conclusions already drawn, showing that employing feature aggregation into metaterms produces better results than the original TF-IDF feature spaces, and than when using the feature space reduction obtained using LSI.

Finally, we show a few examples of the most relevant features within each of the extracted clusters in Table 2. For experiment 1, we show some of the single features which have the highest TF within the cluster, and for experiment 3 we display the metaterms with the highest values. We observe that there are still some features that do not add relevant



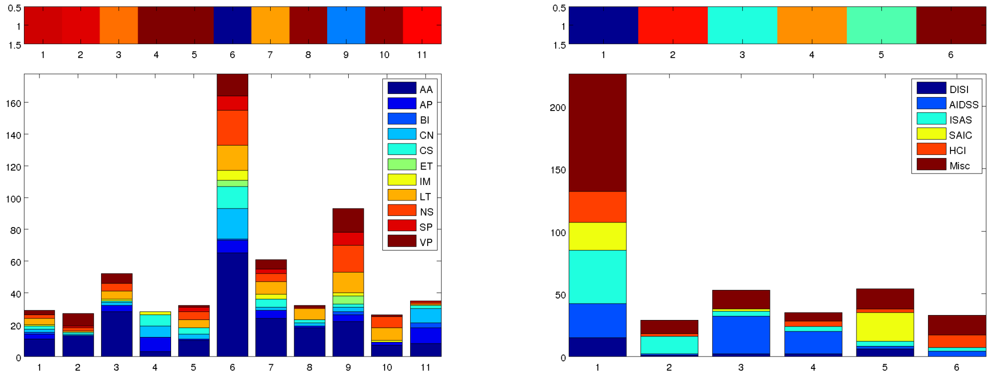
(a) NIPS (b) ICEIS

Figure 5: Graphical representation of the co-association matrices obtained for experiment 1 over the NIPS and ICEIS data set (with document distribution).



(a) NIPS (b) ICEIS

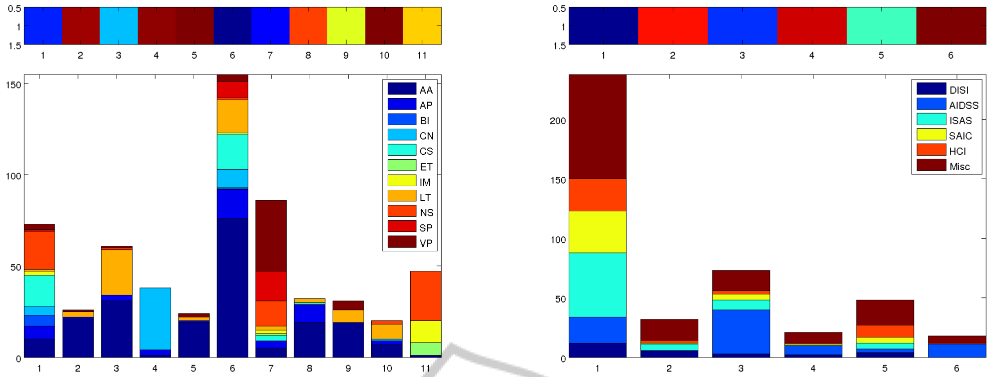
Figure 6: Experiment 1 Results - clustering of documents based on TFIDF - for each cluster we show the distribution of documents by topics (different colors). At the top is represented the confidence of each clustering.



(a) NIPS (b) ICEIS

Figure 7: Experiment 2 Results - clustering of documents based on TFIDF - for each cluster we show the distribution of documents by topics (different colors). At the top is represented the confidence of each clustering.

information to the documents' characterization, such as "summary" or "notable". This suggests that the contextual stopwords-removal step might still be further improved. Regarding the metaterms, in some situations the number of aggregated terms is high (>30), joining terms that can have multiple meanings, introducing errors in the representation.



(a) NIPS

(b) ICEIS

Figure 8: Experiment 3 Results (using Lesk Algorithm) - clustering of documents based on TFIDF - for each cluster we show the distribution of documents by topics (different colors). At the top is represented the confidence of each clustering.

Table 2: Examples of the most relevant features found for some of the extracted clusters. The keywords that serve as root for the metaterms are bolded. Notice how the two aggregation criteria generate aggregate different features based on the same keyword (image processing).

Data Set	Exp.	Cluster Index	Relevant Features
NIPS	1	2	frey; epochs; demonstrates; fischer; decoupled; freedman; dependency; cal; eter; book
		10	calls; centre; dec; extreme; broomhead; avg; disturbance; corporation; affine; colinear
ICEIS	1	3	information;design; systems; paper;location; information systems; context; architecture; method
		4	data; spatial; information; schema; warehouse; mining; query; emergency
NIPS	3 (Lesk)	6	**image processing**; res shape; res shape; images required; ground roc; operation stereo; treat; treats; ground images; images shape; ring; estimates stereo; estimates stereo; proposed shape; sets; experts images; forms model; shape variation; images material; performance shape; images truth; double profiles; implement res; sorts; effect textures; images occlusion; imaging; captures shape; direct shape; variation; model truth; shape stereo; occlusions res; map shape; images texture
NIPS	3 (CoOccur.)	3	**image processing**; instances; temporally; chance; effort; considered; aligned; interface; kernels; multiplied; cropped ;include ;items ;shape shape ;sorted ;specifications ;formalize ;identically ;improved ;kai ;probabilistic ;math ;post ;contaminated ;rows ;consumption ;dendritic ;extend ;joachims ;recipes shape ;arguments ;complexity ;corner ;defer ;designer ;failed ;mika ;notable ;presynaptic ;states ;summary ;terrence ;tradeoff ;tuning
ICEIS	3		**queries** ;query rewriting ;query ; **coresparql**; sparql; **queries**; optimization; optimization query; query optimization; query;

6 CONCLUSIONS AND FUTURE WORK

We proposed a methodology for unsupervised organisation of documents, and in particular research papers, into meaningful groups. The clustering was based on a ensemble approach - Evidence Accumulation Clustering (EAC) - which combines the results of different clusterings, the clustering ensemble. We compared two different documents representations: the typical vector-space-model, and an alternative representation based on metaterms - which are a subset of words and compound terms that are either synonyms or have some kind of semantic relation. Both representations relied on a first step of statistical feature reduction. For the metaterm extraction we devised a criterion based on an adaptation of the Lesk Algorithm which, from keywords or topics assigned to the documents, aggregates words and compound terms (bigrams and trigrams) extracted

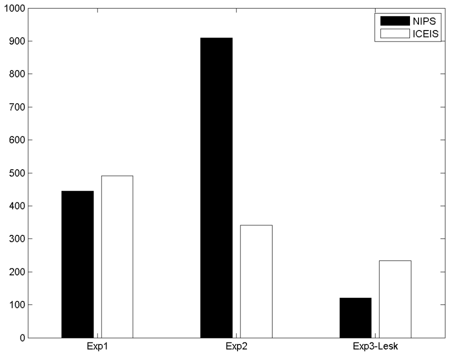


Figure 9: Textual Coherence over different experiments.

from the text. To evaluate the proposed methodology we used two real-word data sets from conferences NIPS and ICEIS. We also evaluated this methodology against results obtained by applying LSI to the original feature space.

To evaluate the results, we followed an unsupervised approach, based on the observation of the obtained co-association matrices, and on the within cluster textual coherence. Based on both, we conclude that feature reduction by employing feature aggregation into metaterms produces better results than both the original TF-IDF feature spaces and the one using the feature space reduction obtained by LSI.

As future work we want to improve the criteria for feature aggregation, including a supervised step of user annotation, and combining different criteria (statistical and string comparison). Additionally, we will use the EAC clustering combination algorithm to combine the information already in use (titles and abstracts) with citation information. Another of the possible approaches is the usage of other ontologies (besides WordNet) for the discovery of semantic relationships between features and documents, enabling better aggregation of features.

ACKNOWLEDGEMENTS

This work was partially developed under the grants SFRH/PROTEC/49512/2009 and PTDC/EIACCO/103230/2008 (project EvaClue) from Fundação para a Ciência e Tecnologia(FCT), and project RETE, reference 3-CP-IPS-3-2009, from IPS/INSTICC, whose support the authors gratefully acknowledge.

REFERENCES

(1998). Acm computing classification system. http://www.acm.org/about/class/1998.

Ahlgren, P. and Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics*, 76:273–290. 10.1007/s11192-007-1935-1.

Aljaber, B., Stokes, N., Bailey, J., and Pei, J. (2010). Document clustering of scientific texts using citation contexts. *Inf. Retr.*, 13:101–131.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810.

Boyack, K. W. and Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., and Brner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE*, 6(3):e18029.

Dao, T. N. and Simpson, T. (2005). Measuring similarity between sentences. http://opensvn.csie.org/WordNetDotNet/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf.

Fellbaum, C. (1998). *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

Fred, A. (2001). Finding consistent clusters in data partitions. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 2096, pages 309–318. Springer.

Fred, A. and Jain, A. K. (2005). Combining multiple clustering using evidence accumulation. *IEEE Trans Pattern Analysis and Machine Intelligence*, 27(6):835–850.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. (2007). Euclidean Embedding of Co-occurrence Data. *The Journal of Machine Learning Research*, 8:2265–2295.

Hanan, G. A. and Mohamed, S. K. (2008). Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):160–173.

Hotho, A., Staab, S., and Stumme, G. (2003). Wordnet improves text document clustering. In *In Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544.

Janssens, F., Leta, J., Glanzel, W., and De Moor, B. (2006). Towards mapping library and information science. *Inf. Process. Manage.*, 42:1614–1642.

Karypis, G., Kumar, V., and Kumar, V. (1998). Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing*, 48:96–129.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, 32:67–71.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documen-*

tation, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Lourenço, A., Fred, A., and Jain, A. K. (2010). On the scalability of evidence accumulation clustering. In *ICPR*, Istanbul Turkey.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Reforgiato Recupero, D. (2007). A new unsupervised method for document clustering by using wordnet lexical and conceptual relations. *Information Retrieval*, 10:563–579. 10.1007/s10791-007-9035-7.

Sebastiani, F. (2005). Text categorization. In *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press.

Sedding, J. and Kazakov, D. (2004). Wordnet-based text document clustering. In *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*, ROMAND '04, pages 104–113, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sevillano, X., Cobo, G., Al?as, F., Socor?, J. C., Arquitectura, E., and Salle, L. (2009). Robust document clustering by exploiting feature diversity in cluster ensembles.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J. of Machine Learning Research 3*.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth, London.

Zheng, H.-T., Borchert, C., and Kim, H.-G. (2009a). Exploiting corpus-related ontologies for conceptualizing document corpora. *J. Am. Soc. Inf. Sci. Technol.*, 60:2287–2299.

Zheng, H.-T., Kang, B.-Y., and Kim, H.-G. (2009b). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, 179(13):2249 – 2262. Special Section on High Order Fuzzy Sets.