An Hybrid Approach for Improving Word Sense Disambiguation and Text Clustering

Paolo Casoto
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 - Loc. Rizzi
Udine, Italy
Email: casoto@dimi.uniud.it

Carlo Tasso
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 - Loc. Rizzi
Udine, Italy
Email: tasso@dimi.uniud.it

Abstract— **In this paper we suggest a new approach to represent text document collections, integrating background knowledge to improve clustering effectiveness. Background knowledge is inferred from previous classification tasks on the same collection and it's used as context to assign semantic values to words. The WordNet ontology acts as a source repository of semantic meanings. More specifically we propose a new approach to Word Semantic Disambiguation, based on the evaluation of functions related with syntactical and statistical properties of the WordNet synsets or with contextual and synonymical informations.**

I. Introduction

Text document clustering can be defined as grouping of documents referring to semantically similar concepts; Hayes defines clustering as the identification of documents that "are grouped because they are likely to be wanted together" [1].

Text document clustering is an autonomous process; it does not need any training activity involving human teachers and the knowledge used during the process comes from the documents' text itself: there is no influence from external knowledge. In accordance with this characteristic, clustering can be considered as an unsupervised learning technique.

On the other hand *text document classification* is based on the experience of a human classifier, which trains properly the system introducing his own subjective knowledge: document classification is then considered a supervised learning technique. An exhaustive comparison between these two processes can be found in [2].

Text document clustering can be a useful advice in Information retrieval activities; Van Rijsbergen [3] describes how cluster hypothesis could be used to improve query reformulation and identification of documents relevant with respect to user informational needs.

In the last few years text document clustering has been used in many new ways, especially as a support for large collections browsing and for navigation of search results. The ever growing size of the World Wide Web [4] and digital libraries can suggest the use of text document clustering in order to implement structured or hierarchical navigation. Interesting examples of this approach can be found in Vivisimo[1] and Carrot[2].

These new clustering approaches carry new requirements, related with both effectiveness and efficiency of the cluster identification process. More specifically, identified clusters must be characterized by an high level of relevance for the user, which means that they must be as similar as possible to the interest model representing the user informational need.

II. Improving the clustering process by means of external knowledge

Many research papers investigated possible ways to improve the performance of text document clustering in terms of efficiency, trying to develop real time clustering algorithms. At the same time there are many unsolved problems related to the improvement of clustering effectiveness [3].

A *label* is a set of one or more words used in document classification to identify a specific class; clustering quality can then be defined as a measure of similarity between result clusters and labels, obtained by comparing the number of documents grouped into a specific cluster with the number of documents assigned to a specific label.

Two factors may influence the cluster's quality:

1) text document clustering does not embodies any external knowledge;
2) unlike humans, computer-based system cannot fully understand the semantics of the natural language text contained in each document of the collection.

These two factors are strongly related to the concept of similarity used in text document clustering. *Similarity* is obtained by an explicit formal evaluation of a predefined set of features of the documents; often the frequency of relevant words is considered the main feature in the text document similarity's evaluation.

However, during document classification, documents may be assigned by the human classifier to a class on the basis of subjective concepts, which are not necessarily related to some explicit feature. For example two documents that share the same relevant words may be assigned to different classes or, on the other hand, not similar documents may be associated with

[1] http://www.vivisimo.com

[2] www.carrot2.org

the same label. Clustering is not able to identify the relations between documents that are not directly bound to a set of explicit features used in similarity evaluation.

The second factor can be seen as a more general problem that arises every time a computer-based system is called to operate with natural language texts. Computer-based systems are still unable to adequately extract the right meaning from a raw text written in natural language. Often the semantics of documents are inferred or guessed from assumptions on the word frequency in the collection. These assumptions may not always be correct or adequate for find the right meaning of a term; for example they are particularly weak when words show phenomena like polisemy or synonymity.

In order to reduce the effects generated by the described factors, two kinds of complementary solutions can be considered:

1) to integrate into the clustering process knowledge generated by supervised learning approaches;
2) to assign to every word that appears in the collection a specific semantic value that depends on word's meaning and context.

The first solution leads to an hybrid approach to document representation, where unsupervised knowledge, implicit in the body of documents, and supervised knowledge are melt together. The main goal of such hybrid approach is to enrich text document clustering with external knowledge and to identify clusters more similar to the classes introduced by a human classifier.

The hybrid approach introduces a new idea of document similarity; more specifically two different kinds of features must be considered during evaluation of similarity between documents: the ones related with documents text and, on the other hand, the ones related with the taxonomy used during the classification process.

In accordance with the second solution proposed, by moving towards a more semantic representation, it is possible to define a new similarity measure between documents, which is based on weights assigned to concepts instead of weights assigned to words. The hybrid approach can be used to identify set of words which act as contexts in Word Semantic Disambiguation process.

The main goal of this research is to design and build a hybrid approach to clustering able to melt unsupervised knowledge with supervised knowledge, which is based on labels assigned to documents as result of a previous document classification activity on the same collection. Labels help to rebuild user taxonomy, which can enrich the results of the clustering process allowing the identification of better clusters.

Moreover we are interested in studying the relationship between the amount of supervised knowledge introduced in the document representation model and the quality of the identified clusters.

The assignment of semantic values to the collection's words is performed by integrating the clustering engine with the semantic knowledge base of WordNet ontology [5].

III. Previous Work

This work is closely related to the work of Wermter *et al.* [6] on hybrid models; it can be seen as an improvement of it. More specifically we suggest a new approach to sub-models harmonization, a clustering algorithm based on a partitioning approach, and a new criterion for word sense disambiguation, in order to increase the quality of generated clusters in terms of similarity with the results of previous document classification processes.

However, while Wermter has focused his attention on Self Organizing Maps (SOM) based clustering, according to Kohonen [7] ideas, in this work we propose to use an optimized Spherical k-Means algorithm [8], which does not require any training task and can manage faster large size vectors.

In the last few years other researchers dealt with the problem of semantic disambiguation of words; in particular Hotho *et al.* [9] and Sedding [10] proposed distinct approaches, based on the integration of background knowledge in the clustering process in order to assign the WordNet concept to words. The necessary knowledge can be extracted from previous classification tasks, like in our work, or from syntactic annotation obtained from supervised taggers. Our suggested criterion for word semantic disambiguation shows better performance compared to previously proposed techniques, in terms of number of right meaning associated to words.

In order to obtain results, which could be compared, we have based our evaluation on the same metrics used in [10].

IV. Model's description

Our hybrid model includes two sub-models, whose goal is to represent, respectively, documents' text implicit knowledge and knowledge inferred from previous document classification tasks. The organization of our model is shown in figure 1.

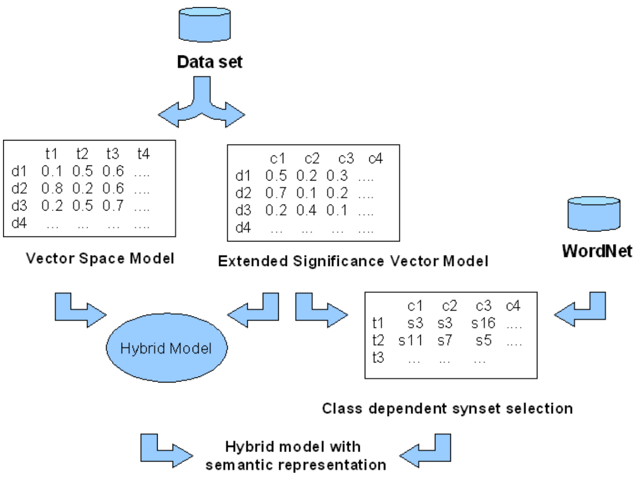
The first sub-model is based on Salton's VSM, which transforms documents into weight vectors; weight are meant to represent the relevance of a specific word in a document. To assign weights to document vectors normalized TF-IDF scheme [11] is used, because it is well suited for large size document vectors. In order to reduce vector size we adopted stopwords filtering, stemming and pruning.

The second sub-model, used to represent inferred knowledge, is based on Wermter's Extended Significance Vector Model (ESVM), which assigns to every document of the collection a weight vector that represents the similarity between the document and every distinct class from a previously defined taxonomy.

In order to reduce ESVM complexity and improve model effectiveness we made two assumptions related to the preclassification process of the documents collection:

1) each document can be assign to at most one class of classification's taxonomy;
2) the taxonomy used during the classification process is a single level taxonomy, which means that there is no hierarchical relation between classes.

The number of taxonomy's classes can be generally considered smaller then the number of distinct terms that appear in the



	t1	t2	t3	t4
d1	0.1	0.5	0.6	....
d2	0.8	0.2	0.6	....
d3	0.2	0.5	0.7	....
d4	...	...	...	....

	c1	c2	c3	c4
d1	0.5	0.2	0.3	....
d2	0.7	0.1	0.2	....
d3	0.2	0.4	0.1	....
d4	...	...	...	....

	c1	c2	c3	c4
t1	s3	s3	s16	....
t2	s11	s7	s5	....
t3	...	...	...	

Fig. 1. Overview of the hybrid model's organization

documents' collection; therefore ESVM's size is not relevant for model efficiency and does not require particular dimensionality reduction tasks or the introduction of optimized data structures. A complete description of the Extended Significance Vector Model can be found in [12].

Sub-models integration into our hybrid model is achieved by building a new document vector for every collection's document; it is formed by concatenating the unsupervised vector based on the Vector Space Model and the supervised vector based on the Extended Significance Vector Model, as described in the following equation:

$$\text{Hybrid vector} = [(1-\gamma) \times \text{VSM vector}][\gamma \times \text{ESVM vector}], \tag{1}$$

where γ is used as a control parameter, with values between 0 and 1. The γ parameter is used to define the influence of each model; when γ is larger, the effect of the supervised part of document vector is more significant. In our work we studied how the γ parameter's values can influence the quality of generated clusters and how to detect a value that cannot only work as a control parameter but also harmonize vector generated by sub-models.

More specifically, although both sub-models represent documents as weighted vectors, these vectors carry different kinds of information, not comparable with each other, and are characterized by a different size. Thus the assignment of values to γ must take care of this dimensional and conceptual differences between the vectors of the two models we are trying to melt together.

To move towards a deeper semantic representation of documents we have exploited supervised knowledge as context for word semantic disambiguation. Word semantic disambiguation allows to assign to each word a specific semantic element picked up from WordNet, called synset. A *synset* can be seen as a set of synonymous terms with a specific id that refers to a specific meaning.

Synset association is guided by the following assumption: "two documents in the same taxonomy's class assign to the same word a unique meaning". Association is based on ESVM: relevant words for each taxonomy's class are used as local context to select, for each word the synset that best matches the meaning of a specific word included in documents belonging to the same class.

Figure 1 can help to illustrate this approach. ESVM sub-model is used to build a representation of relevant words for each taxonomy's class; the relevance of a word for a specific class can be defined as the average relevance of the word for documents assigned to that class. This definition is based on the assumption that a word is relevant in respect with a specific class if it can be used to distinguish between document assigned to the class from the ones not assigned.

The sets of relevant words for a given taxonomy can be represented using a matrix, the Class-Word matrix, which can be seen as the context of word disambiguation process. Class-Word matrix is used by the disambiguation criterion during evaluation of weights to assign to associations between words and synsets.

The disambiguation process can be influenced by two factors:

1) the set of pre-classified documents considered for each class;
2) the threshold used to identify relevant words for a specific class.

Based on Class-Word matrix we propose a new innovative criterion to remove semantic ambiguity and to assign the best WordNet's synset to every word. Our new approach is based on the assumption that four different aspects may be considered during the word semantic disambiguation process.

The first aspect considered by our criterion is related to the syntactical role of the term in the language; we assign an higher rank to those synsets that are related to nouns, because they carry most part of a document meaning. This aspect was previously used in Sedding's work as stand alone criterion to synset selection [10]. WordNet's synset are divided into four syntactical roles: nouns, verbs, adjectives an adverbs; we assign weight 1.0 to nouns, 0.75 to verbs, 0.50 to adjectives and adverbs. With respect to Sedding's work, we do not include in our system a Part-of-Speech Tagging task, in order to identify syntactic roles of the collection's words; the roles are seen as a property of the synsets related with terms.

The second aspect is related to frequency of synsets in the language; we assign an higher rank to those synsets whose frequency is higher, to reduce selection of unusual synsets. Solutions based on synset frequency were previously suggested by Hotho and Brezeale [13]. In particular Brezeale suggested to solve the synset selection problem by always choosing as best synset the most frequent; this kind of approach does not resolve effectively the problem of polisemy and is not useful by itself as a disambiguation criterion, because all occurencies of a given word in the input collection are bounded with a unique synset.

However it may be a useful advice when many similar synsets may be assigned to a term, for example a verb with many different meaning similar each other; selection based on synset frequency may reduce the fan of synsets assigned to term in different contexts. The identification of too large or too small fans of synsets for a given term limits every disambiguation criteria. Wermter and Sedding proposed an alternative solution based on WordNet's hyperonimy relation: synsets are substituted, if possible, by their common parents, reducing the number of synsets used in representation.

The third aspect of our approach is based on symilarity between term's context, defined as a row of the Class-Word matrix introduced previously, and synsets' description. The *description*, known also as *gloss*, of a synset is a short text written in natural language used to define the meaning of the synset, in according with terms belonging to it. Description represents the right use of terms for a given semantic meaning; a measure of similarity between description and term's context can be used to evaluate the goodness of a synset in respect with a specific context. The description's similarity is defined as the cosine similarity between context and synset's description, represented as a vector of weighted terms.

The synset's description may include some phrases representing common use cases of the synset's terms; in this work we don not consider them because they may generate noise during the disambiguation process. A good example of this phenomena is represented by some descriptions associated with the synstes referring to the verb *"see"*.

1) "We found Republicans winning the offices"';
2) "The 1960's saw the rebellion of the younger generation against established traditions";
3) "We went to see the Eiffel Tower in the morning";
4) "Catch a show on Broadway".

Descriptions contain terms with particular meanings, like *Republicans*, *rebellion*, *Eiffel* or *Broadway*, which may be relevant for a given context, and therefore increase cosine similarity between description and context, but not directly related with the meaning of the *"see"* verb.

The fourth aspect considered by our approach is related with the number of synonymous of a given synset that appear in the term's context. A synset can be seen as a data structure based on a set of synonymous terms; a similarity's measure between a context and a synset can be defined as the sum of relevance weights assigned to each synset's synonymous terms in the ESVM representation of the context. The evaluation of similarity between the context and the set of synonymous terms can be very useful to reduce effects generated by synonymity and, at the same time, to improve the quality of the disambiguation process for synset representing rare or very specific meanings.

The four aspect of our approach, described previously, are melted accordingly with the equation 2:

$$SS(S_i, C_j) = \frac{\Gamma(S_i)}{Ranking(S_i)} \times (\alpha\Delta(S_i, C_j) + (1-\alpha)\Phi(S_i, C_j)), \tag{2}$$

where functions $\Gamma(S_i)$,$Ranking(S_i)$,$\Phi(S_i, C_j)$ and $\Delta(S_i, C_j)$ are used to evaluate, respectively, the syntactical role, the frequency, the description similarity and the synonymous terms similarity of a synset S_i. The α parameter is used as control parameter, to allow criterion tuning in case of use on lexically specific collections.

An example may be useful to understand how our new approach works and how the disambiguation task takes place; let *M* be the Class-Word representation obtained from the ESV Model, considering only two different classes and six terms.

$$M = \begin{pmatrix} 0.1 & 0.3 & 0.7 & 0 & 1 & 0.6 \\ 0.4 & 0 & 0.1 & 1 & 0.6 & 0 \end{pmatrix} \tag{3}$$

Let *t* be the term subject to word sense disambiguation; using *t* as key, the following three synsets are found in the WordNet knowledge base:

Index	Sync. Role	Freq.	Terms in gloss	Synonymous Terms
1	noun	1	$\{2,3,5\}$	$\{3\}$
2	noun	2	$\{2,4,6\}$	$\{2,3\}$
3	verb	1	$\{3,4,6\}$	$\{4,5,6\}$

Applying our disambiguation criterion to the selected synsets we obtain the following results:

	Class "a"	Class "b"
Synset 1	0.685	0.225
Synset 2	0.238	0.275
Synset 3	0.544	0.506

The WSD task for term *t* ends with the identification of the best synset to associate to *t* for each different class context; in particular for class *"a"* the synset which best represents term's meaning is synset 1, while synset 3 is the one that best matches the term t in its occurencies in documents assigned to class *"b"*.

The hybrid model obtained after semantic association can be used as input for the clustering algorithm; in our system we used as clustering algorithm an optimized variant of Spherical k-means algorithm, suggested in [8]. This approach is very different from the one adopted by Wermter, based on SOMs; Spherical k-means algorithm was chosen because of its low computational complexity (O(n) with n representing the number of documents of the input collection) and its efficiency in large size vectors handling.

V. Results and Evaluation

In this work we used the evaluation metrics introduced by Sedding [10], namely *purity*, *entropy* and *overall similarity*, to analyze clusters quality. Purity and entropy are used to measure the gap between identified clusters and the collection taxonomy generated during a previous document classification process. Overall similarity is independent from pre-annotations and is used to give an idea of the cohesiveness of a cluster, namely the mean similarity of each couple of documents assigned to a specific cluster.

Purity and entropy are based on precision, which represents the number of documents shared between a cluster and a specific class of the pre-classification's taxonomy. Precision can be evaluated as

$$prec(\pi_i, L) = \frac{|\pi_i \cap L|}{|\pi_i|}, \tag{4}$$

where π_i is a cluster and L is a label that identifies a taxonomy's class.

Purity can be derived from precision; more specifically it is defined as the mean of normalized precision evaluated on each generated cluster, as

$$purity = \sum_{\pi_i \in \Pi} \frac{|\pi_i|}{|D|} max_L prec(\pi_i, L), \qquad (5)$$

where $|D|$ is the number of documents of the collection, used for normalization. Purity values can vary in the range between 0 and 1.

Entropy represents the mean normalized intra-cluster amount of disorder, in terms of dispersion of the documents associated with a taxonomy's class between the distinct clusters; more specifically entropy is calculated as:

$$entropy(\Pi) = \sum_{\pi_i \in \Pi} \frac{|\pi_i|}{|D|} \times ice(\pi_i), \qquad (6)$$

where Π is the set of generated clusters and *ice* is an extra function, the intra-cluster entropy, that evaluates the entropy of a single cluster, using the equation:

$$ice(\pi_i) = \sum_{L} prec(\pi_i, L) \times \log(prec(\pi_i, L)). \qquad (7)$$

To evaluate system performance we decided to adopt the corpora defined in [10], extracted from Reuters-21578 text collection. Composed by 21578 articles published in 1987 by Reuters, this corpus has been used to evaluate many IR and clustering systems. The corpus is not domain specific and part of its documents are pre-classified, which means that their labels can be used to enrich the hybrid model.

In order to evaluate our system on a collection of documents with heterogeneous distribution of the pre-defined taxonomy, new sub-corpora, containing only pre-classified documents, have been extracted from the original corpus. Moreover the set of pre-classified documents is filtered, to remove taxonomy classes with unbalanced distribution of documents and documents assigned to more than one class. We obtained four new corpora, called respectively:

1) *Reuters 100*: every class of the classification taxonomy with at most 100 documents (classes with more than 100 documents are reduced in size using a random selection of the documents assigned to them);
2) *Reuters 50*: every class of the classification taxonomy with at most 50 documents;
3) *Reuters 20*: every class of the classification taxonomy with at most 20 documents;
4) *Reuters 20-15*: every class of the classification taxonomy with at most 20 and at least 15 documents(classes with less then 15 documents are removed from the taxonomy).

The goal of evaluation was to find a relation between the amount of supervised knowledge introduced into the hybrid model and the quality of identified clusters, measured in terms of purity, entropy and overall similarity. In addition we wanted to prove that the proposed word semantic disambiguation criterion may increase clustering quality by introducing in the model more semantic information.

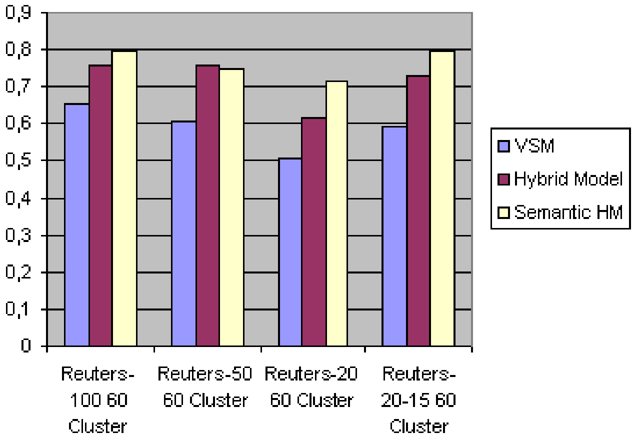


Fig. 2. Purity of clusters generated from different representation models on test corpora

Figure 2 shows the most significant results, obtained by comparing purity of clusters generated from a standard VSM representation, a hybrid model and a hybrid model with semantic representation of documents. The results are obtained using the four test corpora described previously as input for the clustering engine and assigning to the γ parameter value 0,5, which means that same importance is given, in the hybrid models, to both VSM and ESVM representations.

The hybrid model improves significantly the purity of identified clusters (reducing at same time clusters entropy), which results to be more similar to human taxonomy, thanks to the external knowledge included in the clustering process. The significant improvement's rate between VSM representation and hybrid models can be considered the same for all the test corpora; in particular it shows how background knowledge can really improve clustering effectiveness in terms of purity end entropy of identified clusters.

On the other hand the result show how semantic representation may introduce different improvement into hybrid model; in the second of our test corpora, shown in figure 2, the hybrid model with no semantic representation performs better than the one with the semantic representation.

To evaluate the quality of our word semantic disambiguation approach we compared it with the one proposed by Wermter based on similarity between term context and synset's gloss. To analyze the effects of the new approach we evaluated not just the quality of identified clusters on the four test corpora, but also the size of the fan of synsets assigned by both methods to collection's terms.

The diagrams in figure 3 and 4 show how our approach leads to a relevant reduction in the number of synsets assigned to collection's thesaurus respect with the Wermter's approach. More specifically it is relevant to see how the proposed dis-

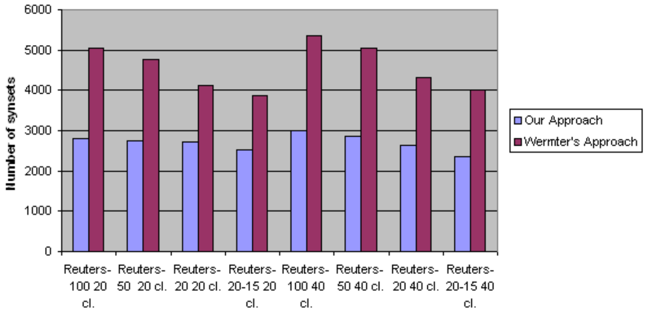


Fig. 3. Total number of synsets identified by the two disambiguation approaches

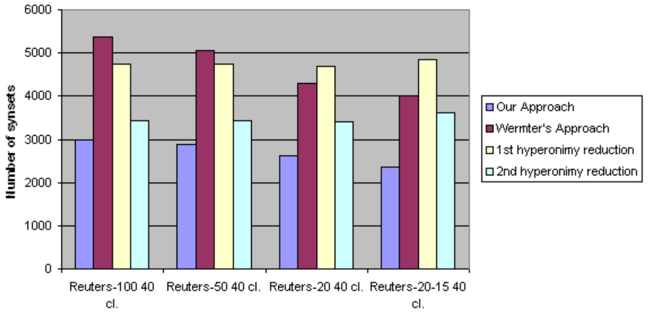


Fig. 4. Comparison between synsets identified by both methods and one/two steps ”hyperonimy based” reduction

ambiguation method acts as a filter better than the hyperonimy relation, used by Wermter and Sedding as bound to growth of synsets number. The figure 4 extends figure 3 introducing two more elements, the number of synsets identified by Wermter approach after one or two steps of the ”hyperonimy based” reduction. A consequence of this reduction in the size of the synsets' fan is a reduced size of the semantic hybrid model and an improvement of model performance.

The improvement introduced to the semantic hybrid model by our new approach is significant and can be seen in figure 5, where is shown the purity of clusters identified applying both semantic models to the four test corpora.

VI. Conclusions and future work

Our work shows how an hybrid model can improve clustering effectiveness in terms of purity respect to human taxonomies. In addition it shows how our word semantic disambiguation criterion can help transition towards a more semantic representation of documents, selecting the best synset for a given word in a specific context represented by an external taxonomy. With respect to other criteria described in previous works, our approach shows a better performance, when applied to the chosen test collections.

The clustering engine we developed may have many interesting applications in the area of digital libraries: it can be used as an advice in document browsing, grouping together similar documents and, consequently, reducing complexity of collection's navigation tasks. Moreover clustering can be exploited as an aid to support manual indexing and classification.

Future research goals may include the integration with the italian version of WordNet's ontology and a deeper investigation of the relationships between the specific supervised knowledge introduced into the model and the quality of generated clusters.

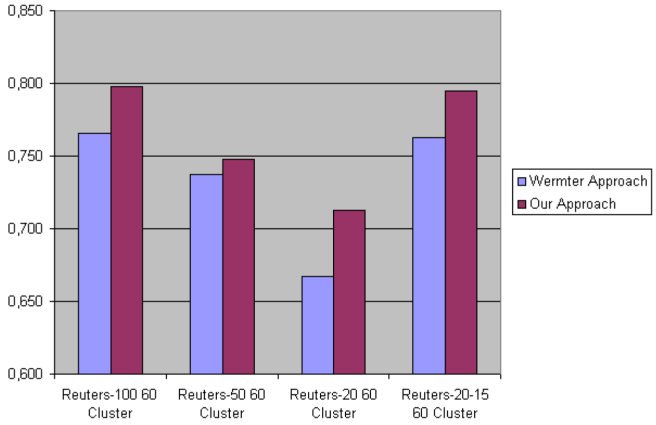


Fig. 5. Purity of identified clusters

References

[1] R. Hayes, *Mathematical models in information retrieval*, 1963, p. 287.

[2] M. Hearst, *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999, pp. 333–374.

[3] C. van Rijsbergen, *Information Retrieval 2nd Edition*, 1989.

[4] A. S. A. Gulli, “The indexable web is more than 11.5 billion pages,” in *WWW 2005*. [Online]. Available: citeseer.ist.psu.edu/gulli05indexable.html

[5] G. Miller, “Wordnet: An on-line lexical database,” *International Journal of Lexicography*, vol. 3(4), pp. 235–244, 1990.

[6] P. S. C. Hung, S. Wermter, “Hybrid neural document clustering using guided self-organisation and wordnet,” *Issue of IEEE Intelligent Systems*, vol. March/April, pp. 68–77, 2004.

[7] T. Kohonen, “Self.organized formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[8] Y. G. I. Dhillon, J. Fan, “Efficient clustering of very large document collections,” in *Data Mining for Scientific and Engineering Applications*, R. N. R. Grossman, G. Kamath, Ed. Kluwer Academic Publishers, 2001. [Online]. Available: citeseer.ist.psu.edu/dhillon01efficient.html

[9] G. S. A. Hotho, S. Staab, “Wordnet improves text document clustering,” 2003. [Online]. Available: citeseer.ist.psu.edu/hotho03wordnet.html

[10] J. Sedding, “Wordnet-based text document clustering,” Master's thesis, University of York, 2004.

[11] G. Salton, *Automatic text processing: the Transformations, Analysis, and Retrieval of Information by Computer*. Addison Wesley, 1989.

[12] P. S. C. Hung, S. Wermter, “Predictive top-down knowledge improves neural exploratory bottom-up clustering.” [Online]. Available: citeseer.ist.psu.edu/631988.html

[13] D. Brezeale, “The organization of internet web pages using wordnet,” Master's thesis, University of Texas at Arlington, 1999. [Online]. Available: citeseer.ist.psu.edu/brezeale99organization.html