Content-Based News Retrieval on the Web

Pasquale Capasso, Carmine Cesarano, Antonio Picariello, Lucio Sansone

Dipartimento di Informatica e Sistemistica, University of Naples "Federico II", Italy

Summary
Effective search and retrieval are fundamental for realizing the full potential of the Web. Although nowadays search engines perform much better than few years ago, big improvements are still needed with respect to the relevance of the retrieved documents to the user's query and the presentation of the results. In this paper we present a novel retrieval system which exploits Wordnet's semantics in identifying the topic of News documents and ranks the retrieved ones according to their relevance to the query. Furthermore, the system provides the user with a short summary of each document, helping her/him in browsing the result set.

Key words:
Information Retrieval, Relevant Terms extraction, NPL, clustering.

Introduction

The Web is nowadays the biggest and most various existing knowledge base, providing users with huge amount of data in disparate fields. The data spread out on the Net, the newspapers, the on-line magazine and in general the web are, in fact, of strategic importance to a plethora of applications, such as business, defense agencies, etc.. On the other hand, the size and disorganization of these data poses great problems for what concerns access to and retrieval of relevant documents, greatly complicating the task of browsing the information retrieved by search engines. In latest years many efforts have been made to develop systems act to simplify the browsing through retrieved documents. Most recent approaches mainly rely on exploiting the semantics of keywords provided by the user as input query [1], [6] with or without the help of external knowledge bases (e.g., Wordnet or some specialized ontology) or user feedback. Other approaches instead rely on the identification and modeling of topics in retrieved documents and the design of opportune clusters, devoting most efforts towards a better organization of the information [2], [4], [3], [8].

In this paper, we describe how to develop a retrieval system that can perform searches onto a local news database, where stored documents are retrieved from the Web, analyzed by a syntactic-semantic parser and clustered based on the characterizing topics. The input query consists of an example document, based on which the system can infer the topics of interest, with an approach similar to the one followed by modern Topic Detection systems [6], [5], [7], [9], [13], [14].

The system then tries to match the topics extracted from the query example and the topics labeling the documents in the local knowledge base, which is continuously updated by the crawler retrieving news from the Net. When a match occurs, the presentation module shows the ranked list of relevant documents.

2. Theoretical Approach

This section describes a novel approach to retrieve documents in according to the TDT paradigm. To this aim a set of features are extracted from a given document, a clusterization is performed and a retrieval process is defined.

2.1 Feature extraction

From a general point of view, we characterize a document on the basis of the well known "4W (Who, Where, When, What) paradigm" (see Makkonen et al. [4]), i.e. we extract the following words from the text:

- Names, of people or organizations (Who);
- Geographical Locations (Where);
- Dates (When);
- Nouns, verbs and/or adjectives describing the events presented in the document (What).

As broadly described in the literature, the extraction of these words may be performed by a *tagger* which discerns named-entities from common Part Of Speech (POS) words, eliminating stop-words with the help of a dictionary. A *Named Entity Detector* is used to extract the list of Persons, Organizations and Dates; the list of common nouns that most characterize the content of a document is determined considering the semantics of each POS word by exploiting the relationships provided by WordNet.

Initially all the nouns contained in the document are candidate to represent the topics of the document. Then we expand each noun considering its list of synonyms (as provided by Wordnet) and a two level hypernym and hyponym expansion. For each word the set of meronyms is also considered. The aim of such an expansion is building a Semantic Hierarchy (SH) upon each given noun (see definition (1)).

The choice of expanding the nouns with only two levels of hypernyms and hyponyms reflects the need of individuating terms which are close enough to be easily related, avoiding at the same time false linkages between semantically uncorrelated words (something that can easily happen given the connected nature of WordNet).

$$SH(w_i) \equiv \left\{ \begin{array}{l} s_i, h^1_{\{0,1\}}, \ldots, h^1_{\{0,m\}}, h^2_{\{1,1\}}, \ldots, \\ h^2_{\{1,n\}}, \ldots, h^2_{\{m,1\}}, \ldots, h^2_{\{m,p\}}, \\ m_i, m^1_{\{0,1\}}, \ldots, m^2_{\{m,p\}}, hy^1_{\{0,1\}}, \\ \ldots, hy^1_{\{0,m\}}, hy^2_{\{1,1\}}, \ldots, hy^2_{\{m,r\}}, \\ my^1_{\{0,1\}}, \ldots, my^2_{\{m,r\}} \end{array} \right\} \tag{1}$$

where: s_i is the set of the synsets associated to the word w_i; $h^j_{\{k,l\}}$ is the hypernym of j-th level (w.r.t. the root of the hierarchy) of the l-th synset associated to the k-th noun; m_i is the set of meronyms associated to s_i; $m^j_{\{k,l\}}$ is the set of meronyms associated to $h^j_{\{k,l\}}$; $hy^j_{\{k,l\}}$ is the hyponym of j-th level of the l-th synset associated to the k-th noun; and $my^j_{\{k,l\}}$ is the set of meronyms associated to $hy^j_{\{k,l\}}$.

Observe that two different root words (i.e., the nouns on which two different SH are built upon) could be considered semantically related if any of the superior or inferior nodes share the same meronym set (see example in fig.1); thus we match the Semantic Hierarchies looking only for the intersection of the meronym sets. The set of candidate terms is then restricted solely to those words whose Semantic hierarchies intersect.

Hence, considering two words w_i, w_j, these words are candidate terms iff:

$$\upsilon_i \cap \upsilon_j \neq \{\phi\}$$

being

$$\upsilon_i = \{s_i, m_i, m^1_{\{0,1\}}, \ldots, m^2_{\{m,p\}}, my^1_{\{0,1\}}, \ldots, my^2_{\{m,q\}}\},$$
$$\upsilon_j = \{s_j, m_j, m^1_{\{0,1\}}, \ldots, m^2_{\{n,r\}}, my^1_{\{0,1\}}, \ldots, my^2_{\{n,s\}}\},$$

In the example in fig.1, given two nouns {scooter, camion}, we build the associated hierarchies according to definition (1). For sake of simplicity, only two branches for each hierarchy are depicted and only the hypernym expansion is considered. The bottom circles represent the roots starting from which two consecutive hypernym and hyponym expansions are applied, completing the vertical growth of the hierarchies. In the final stage we apply the meronym expansion to each node. In figure 1 the set of meronyms of each node is shown in a circle and the intersecting sets belonging to different hierarchies share the same color.

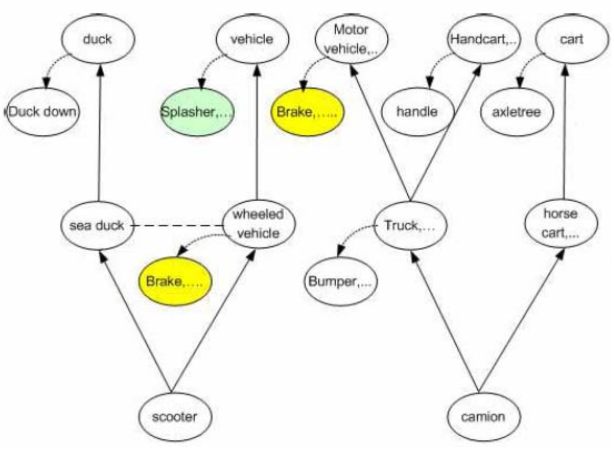


Fig.1. Semantic Hierarchies built upon the words *scooter* and *camion*

As previously anticipated, we match the meronym sets to avoid that two nouns belonging to the same conceptual domain (but with no hypernyms/hyponyms in common in the first two levels) would not be linked (e.g. the nouns scooter and camion do not have common hypernyms/hyponyms, even if belonging to the same "vehicle" domain). Using this approach it is also possible to define the context in which the word is used. In fact there are words having more than one meaning, depending on the context in which they are used.

During the matching step, the Semantic Hierarchies are pruned of the branches whose nouns do not give contribution to the matching. In this way, each hierarchy is restricted solely to the most relevant part.

Considering again the example in fig.1, the first branch depicted (the one containing 'sea duck' and 'duck') is pruned, given that the meronym sets of all nodes in it do not match the meronym set of any node in the second hierarchy.

Given the set of candidate nouns and the relative pruned Semantic Hierarchies, a "*grade of relevance*" is associated to each of these nouns, according to the following formula:

$$P(w_i) = \frac{m \cdot f_0 + p \cdot f_1 + q \cdot f_2}{f} \tag{2}$$

where: f_0 is the number of times that the noun w_i appears as root noun in the set of hierarchies of the candidate selected terms; f_1 is the number of times w_i appears as hypernym or hyponym of the first level; f_2 is the number of times w_i appears as hypernym or hyponym of second level; f is the number of times w_i appears as hypernym or hyponym of first or second level, meronym or root of a hierarchy; *m, n, p* are constants set in such a way that the hypernyms/hyponyms of first level have a greater weight than hypernyms/hyponyms of second level. Based on the maximization of the grade of relevance (2), the top k candidate nouns are chosen as most characterizing the topics of the document.

2.2 Building document clusters

The set of features (i.e. Relevant Terms, Names, Location, Organization, Date) are used to cluster the collection of news document fetched constantly from the different data sources. To build a set of clusters, the first step is to weight each relevant term of a preliminary group of news document, using the following formula:

$$w_{t,n} = \frac{tf_{t,n} \cdot \dfrac{\log(0.5+N)}{df_t}}{\log(1+N)} \qquad (3)$$

where $tf_{t,n}$ is the number of times a relevant term occurs in the news document n considering also all the synonyms of such words, df_t is the number of news document in which the term t occurs considering also all the related synonyms. The weighted relevant terms are organized in a vector that represents the content of the document. In this way long documents are represented by a focused set of terms limiting the storage space for vectors. In order to compute the similarity of two sets of relevant terms belonging to two different news documents, the following formula is applied:

$$\mathrm{Re}\,l(A,B) = \frac{\sum_i \sum_j w_{ti,A} \cdot w_{tj,B} \cdot \sigma(w_{ti,A}, w_{tj,B})}{(\sum_i w_{ti,A}^2 \cdot \sum_j w_{tj,A}^2)^{0.5}} \qquad (4)$$

where $w_{ti,A}$ and $w_{tj,B}$ are the weighted selected terms in the documents A, B respectively and $\sigma(w_{ti,A}, w_{tj,B})$ is a function to determine the similarity between two words. In particular the similarity metrics based on document features proposed by Lin [15] is adopted. The Rel measure is normalized to range in the interval [0,1].

The global semantic similarity between the document A and document B is computed by combining the contribution of the single feature as shown in the following formula:

$$sim(A,B) = \omega_1 \mathrm{Re}\,l(A,B) + \omega_2 Nam(A,B) + \omega_3 Loc(A,B) \ (5)$$

where $\omega_1, \omega_2, \omega_3$ are the weights on the different features. The sum of that weights is equal to 1. Loc(A,B) is 1 if there is some location that appears in both documents, otherwise it is 0. Nam(A,B) is the set of names and organization defined in the same manner of Loc(A,B).

If the similarity to the nearest neighbor is above a threshold theta = 0,34 (determined empirically) assign the news document to the cluster of that neighbor. If the similarity is below the threshold theta, then form a new singleton cluster containing just that news document. We represent every existing cluster by its centroid, that is basically the average of all news documents in the cluster.

In this way the centroid is the sum of all weighted relevant terms present in the news document belonging to the cluster and the collection of all named entity (i.e. Person Name, Organization) and locations. The centroid of a cluster is following represented:

$$\chi = \{Vett_{relevantTe\ \ rms}\} \cup \{NamedEntit\ \ y\} \cup \{Location\ \ \}$$

where $\{Vett_{relevantTerms}\}$ is the set all weighted relevant terms of all the document in the cluster, $\{NamedEntity\}$ is the set of all named Entity and $\{Location\}$ is the set of all locations in the documents.

Each time a new document is added to a given cluster the centroid is update in terms of new relevant terms, named entities and locations.

As a result of this approach, a news document is inserted in a cluster only if it matches, on average, all of the news documents in a cluster more than in any other.

3. Document retrieval

Supposing that a user queries the system by means of an example document (i.e. query by example) and she/he is interested in finding all the documents presents in the repository that are related to the event described in the submitted document, the *query document* is analyzed to extract the set of features in the same way described in the paragraph 3.

To each relevant term is applied the *term frequency - inverse document frequency* paradigm by means of the formula (3). Considering that the collection of documents is represented in the repository by a set of centroids, the retrieval process is performed using the features related to such centroids. In particular, the Rel formula (4) and the similarity formula (5) are applied considering only the relevant terms, named entities and locations of the query document and the relevant terms, named entities and locations contained in the centroids present in the repository.

The set of news document retrieved by the system are all the documents represented by the centroids having the higher score of similarity respect to the query document. The results are presented on the base of news document authored date.

4. System Architecture and algorithms

The system architecture is presented in the figure 2.

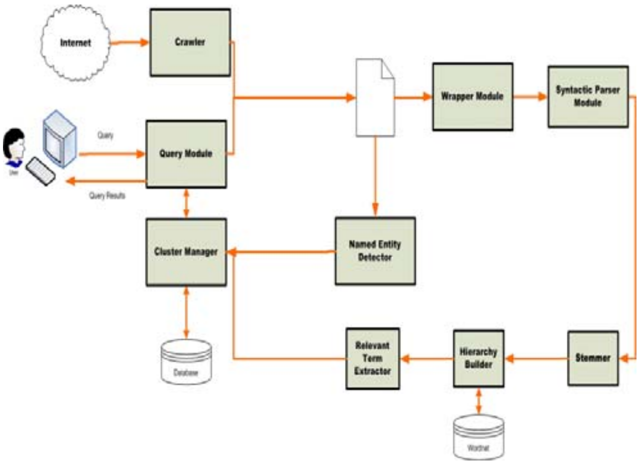


Fig. 2: System Architecture

The Crawler module catches news articles from a set of specified web sources. It reads the RSS files offered by the different websites and every time a RSS file is updated, it fetches the new published documents. A specific *wrapper*, for each web source, has been developed that permits to extrapolate, in an accurate way, the plain text contained in the HTML page. The extracted text is analyzed by the *syntactic module* which identifies the syntactical structure of any single sentence.

The syntactic module extracts only the nouns, which are subject to further elaboration. The choice of considering only nouns is due to the fact that it is commonly argued that language semantics is mostly captured by nouns.

The set of nouns are sent to a *stemmer* that identifies and extracts the core root from those nouns. The stemmed nouns are received by the *Hierarchy Builder*, which, using the relationships provided by WordNet, creates Semantic hierarchies as described in section 2.1. The set of semantic hierarchies are appropriately stored in the database and are used by the *Relevant Term Extractor* to select the list of *relevant terms*, i.e. the nouns which show the highest characterizing power for the considered document.

The *Named Entity Detector* module extracts named entities using a statistical approach based on a standard hidden Markov model as described in [18].

The set of all features are stored in the database and are used during the retrieval process.

The *Cluster Manager* has the task of building and maintaining the cluster as described in the section 3 and, at the same time, it allows to perform the comparison between the query document features and the centroids of the different clusters stored in the database.

The *Query Module* permits to submit queries to the system using a graphical user interface and allows to show the most similar documents to a proposed one.

5. Experimental Results

Our approach differs significantly from conventional information retrieval methods, so there is no standard method to evaluate its performances. Hence, we limited the experiments to the proof of the effectiveness of the extracted features to the representation of the documents' topics. We also evaluated the precision and recall of the clustering and retrieval processes.

5.1 Feature extraction

To evaluate the precision and recall offered by the system during feature extraction, we have enrolled 10 students who were asked to read a collection of documents extracted form "Reuters-21578" and a collection of document extracted from "The 20 Newsgroups" data set. The total number of selected documents was 400, divided in 200 news articles from the first collection and 200 documents extracted from the second data source. We considered two different data sets because of their characteristics; in fact the first one is written in a formal way while the second one contains colloquial and informal expressions. In this way, we wanted to test the proposed approach against different writing styles.

The students had to read the selected documents and underline the terms that better characterize the documents. For this process the students were not instructed on how to select the relevant terms; the choice was completely left to their understanding of the text. For each document there has been a common set of relevant terms identified by all the students while some particular terms were only captured by some of them. To overcome this problem we considered as "identified" terms that were individuated by at least three students. The results of their evaluation have been compared to the features extracted by the system.

The precision and recall are formalized by the following formulas:

$$\Pr ecision = \frac{R}{R+NR} \ , \tag{7}$$

where R is the number of relevant terms retrieved by the system; NR is the number of not-relevant terms retrieved by the system.

$$\mathrm{Re}\, call = \frac{R}{R+RNR} \ , \tag{8}$$

where R is the number of relevant terms retrieved by the system; RNR is the number of relevant terms not retrieved by the system.

A measure of the effectiveness of the system is obtained by means of the *F1* measure that is commonly used to combine precision and recall scores [10]:

$$F1 = \frac{2 \cdot \Pr ecision \cdot \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call}, \qquad (9)$$

where higher *F1* scores (whose values range in the interval [0,1]) are associated to better performances of the system.
The system obtain a maximum precision of 0.82 and a maximum recall of 0.83 (on the first dataset), while the average values on the entire document collection is 0.71 and 0.72 respectively. Some problems are individuated in processing documents belonging to the second document collection. As previously anticipated, the second collection has documents containing informal sentences and a set of words that are very difficult to correlate with the others due to the limitation provided by the use of such a general lexicon as WordNet. As result, the precision and recall on the second data set are approximately 0.54.
A summary of the precision, recall, and *F1* score for relevant terms is provided in the table 1:

Tab. 1 Precision/Recall/F1 for the 2 datasets

Precision	***Recall***	***F1***
0,71	0,72	0,71
0,55	0,53	0,54

5.2 Document Clustering process

To evaluate the effectiveness of the clustering process, in our experiments we use the standard TDT evaluation measure [16]. It is based on a cost function that is a weighted combination of miss and false alarm rates:

$$C_{\det} = C_{Miss} \cdot P_{Miss} \cdot P_{t\arg et} + C_{FA} \cdot P_{FA} \cdot P_{non-t\arg et} \qquad (10)$$

Where C_{Miss} and C_{FA} are user-specified costs for a miss and for a false alarm, respectively; P_{target} is the prior probability that a story will be on topic; $P_{non\text{-}target}$ is the prior probability that a document is not on topic (P_{target}=*1*-$P_{non\text{-}target}$); P_{miss}, P_{FA} are the conditional probabilities of a miss and a false alarm, respectively (i.e., the actual system error rate). To evaluate this cost function, we fixed the cost of a missing detection and the cost of a false alarm to 10 and 1 respectively in according to the TDT evaluation program. A subset of the "Reuters" dataset is used to determine the best combination of the parameters introduced in the formula (5) and the "theta" parameter. The parameters are chosen to maximize the cost function (10). The combination of parameters shown in the table 2 permits to achieve a cost function value of 0,7102 that is a little bit more than the value obtained in [17].

Tab. 2 Parameters values

ω_1	ω_2	ω_3	*theta*
0,45	0,38	0,27	0,34

This combination of parameters permits to cluster together documents that address the same events.

5.3 Retrieval process

As described in section 2, the user has the possibility to query the system with an example document in order to get similar documents (addressing the same issues). Besides, the users can specify the values of the parameters used in the equation (5). The choice of the parameters is influenced by which information the user is interested in. For instance, if she/he wants to retrieve all the documents addressing a specific fact or event, she/he may specify values like 0.4, 0.3, 0.3 respectively in order to weigh up in equal way Relevant Terms, Persons and Locations. Instead, if she/he is interested only in finding documents containing Locations and Persons, ignoring completely the context in which such entities and locations are mentioned, she/he can set the parameter: 0, 0.5, 0.5.
To evaluate the effectiveness of the retrieval process, we initially test the system on an ad hoc dataset, for which we know exactly the number of documents related to specific events, then we test the system on a real data set considering the human judgment to measure the effectiveness of the system.
For the first experiment, we selected 5 random documents as query documents in a database of 1000 documents and, on the basis of the obtained results, we computed the average precision, recall and F1 score.
To run the experiment, we considered that the user is interested in retrieving documents that describe a specific event so the similarity parameters have been set to 0.4, 0.3, 0.3 respectively.
The comparison between the query document and the collection of documents in the database is conducted comparing the features extracted from the query document with the features of the centroid of each clusters. Returned documents are the ones belonging to the cluster with the highest similarity value.
The performances of the system are reported in the following table:

Tab. 3 Precision, Recall, F1 score of the retrieval process

Precision	***Recall***	***F1***
0,84	0,81	0,83

The second experiment aims to evaluate the system on a real context; the crawler module is used to populate the database with about 15000 news documents from news websites. Ten different students were enrolled to judge the precision of the system on five random queries (chosen among the documents in the populated database).

In order to evaluate the recall, we used the Vector Space Model [11]. For each query, the results obtained by our system and the results obtained using the Vector Space Model are merged to form the set of correct results so that a relative judgment of the proposed method is possible [12]. Using the proposed methodology in the following tables the average values of precision, recall and F1 measure are reported.

Tab.4 Precision, Recall, F1 score for a real database

Method	***Precision***	***Recall***	***F1***
SVM	0,52	0,42	0,46
Proposed	0,72	0,63	0,67

The table indicates that the proposed method is more effective than the VSM; in fact it achieves *F1* values of 0.2 more than the state of art model. This result is mostly due to the fact that our approach consider the relevant terms of a document combined with the Named Entity and Location while the SVM uses only the exact matching between words.

6. Conclusions and future works

In this paper we have described a system able to perform searches onto a local database, where stored documents are retrieved from the Web by means of a crawling agent, analyzed by a syntactic-semantic parser and clustered in the database on the base of the characterizing events. We have used, from on one hand, WordNet in order to build suitable algorithms for detecting the terms that more characterize a document and, from the other one, a semantic similarity measure based on cluster centroids to identify documents sharing similar information.

Several experiments have been carried out, in terms of recall and precision in order to evaluate the effectiveness of the features extraction process, the document clusterization and the retrieval one. The results are encouraging so that future works includes experimentation with more data sets like TREC.

References

[1] M.Albanese, P.Capasso, A.Picariello, A.M.Rinaldi: “Information Retrieval from the Web: An Interactive Paradigm”, Proceedings of International Conference on Advances in Multimedia Information Systems (MIS 2005), September 2005.

[2] J.Allan, J.Carbonell, G.Doddington, J.Yamron, Y.Yang: “Topic Detection and Tracking Pilot Study Final Report”, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, February 1998.

[3] J.Carthy: “Lexical Chains for Topic Tracking”, PhD Thesis, Department of Computer Science, National University of Dublin, 2002.

[4] J.Makkonen, H.Ahonen-Myka, M.Salmenkivi: “Applying semantic classes in event detection and tracking”, Proceedings of International Conference on Natural Language Processing (ICON 2002), 1999.

[5] J.Makkonen, H. Ahonen-Myka, M.Salmenkivi: “Simple Semantics in Topic detection and Tracking”, Information Retrieval, Kluwer Publishers, 2004.

[6] R.Mihalcea, S.Mihalcea: “Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web”, International Conference on Tools with Artificial Intelligence, 2001.

[7] P. van Mulbregt, I.Carp, L.Gillick, S.Lowe, J.Yamron: “Text Segmentation and Topic Tracking on Broadcast News via a Hidden Markov Model Approach”, Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998), 1998.

[8] R.Nallapati: “Semantic Language Models for Topic Detection and Tracking”, Proceedings of the HLT-NAACL 2003 Student Research Workshop, 2003.

[9] C.Clifton, R.Cooley, J.Rennie: “TopCat: Data Mining for Topic Identification in a Text Corpus”, IEEE Transactions on Knowledge and Data Engineering, 2004.

[10] Y. Qi, A. Hauptmann, and T. Liu: “Supervised classification for video shot segmentation” in Proc. IEEE Conf. Multimedia Expo (ICME03) vol. 2, 2003, pp. 689-692.

[11] G. Salton. Automatic Text Processing: “The Transformation Analysis and Retrieval of Information by Computer”. Addison-Wesley, 1989.

[12] E. Voorhees and D. Harmann: “Overview of the Seventh Text Retrieval Conference (TREC-7)”. In NIST Special Publication 500-242: The Seventh Text Retrieval Conference (TREC-7), pages 1-23, 1998.

[13] Y.Yang, T.Ault, T.Pierce, C.W.Lattimer: “Improving Text Categorization Methods for Event Tracking”, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in IR, ACM Press, 2000.

[14] Y.Yang, J.Zhang, J.Carbonell, C.Jin: ”Topic-conditioned novelty detection”, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining, 2002.

[15] Dekang Lin: "An information-theoretic definition of similarity", in Proceedings of the 15th International Conf. on Machine Learning, pp. 296-304. Morgan Kaufmann, San Francisco, CA, (1998).

[16] The 2002 topic detection & tracking task definition and evaluation plan. www.nist.gov/speech/tests/tdt2002/evalplan.htm 2002

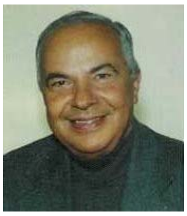
[17] Yang,Y, Pierce,T & Carbonell,J.A A study on retrospective and on-line Event Detection, in proc. of SIGIR 1998

[18] M. Daniel, Bikel and Richard L. Schwartz and Ralph M. Weischedel: "An Algorithm that Learns What's in a Name", Machine Learning Journal, volume 34, pages = 211-231, 1999

Pasquale Capasso received the Laurea degree in Electronic Engineering from the University of Napoli, Italy, in 2004. In the same year he joined the Dipartimento di Informatica e Sistemistica, University of Napoli "Federico II", where he's currently a Ph.D. student in Computer Science and Engineering. His current research interests lie in Information Retrieval and Bioinformatics.

Carmine Cesarano received the Laurea degree in Computer Science and Engineering from the University of Napoli, Italy, in 2002. In 2002 he joined the Dipartimento di Informatica e Sistemistica of the University of Napoli "Federico II", as research follow and at the end of 2003 he started a Ph.D. program in Computer Science and Engineering. He is currently at the last year of his Ph.D. program and his current research interests lie in Information Retrieval , Knowledge Extraction and Management, Natural Language Processing.

Antonio Picariello received the Laurea degree in electronics engineering and the Ph.D. degree in computer science and engineering, both from the University of Naples, Naples, Italy, in 1991 and 1998, respectively. In 1993, he joined the Istituto Ricerca Sui Sistemi Informatici Paralleli, The National Research Council, Naples, Italy. In 1999, he joined the Dipartimento di Informatica e Sistemistica, University of Naples "Federico II," and is currently an Assistant Professor of Data Base and Senior Researcher. He has been active in the field of computer vision, medical image processing and pattern recognition, object-oriented models for image processing, and multimedia database and information retrieval. His current research interests lie in knowledge extraction and management, multimedia integration and image and video databases. Dr. Picariello is a Member of the International Association for Pattern Recognition (IAPR).



Lucio Sansone received a Laurea degree in Electronic Engineering in 1965 from the University of Naples "Federico II". From 1976 to 1980 he has been chair of "Centro di Studio sui Calcolatori Ibridi" (Hybrid Computer Research Center, The National Research Council of Italy, C.N.R.) and he has been Project Leader of C.N.R. research project "Information Systems and Parallel Computing". He joined the Department of Computer Science and Systems of the University of Naples (Italy) in 1965. Since 1980 to present he is a *Full Professor* of Information Systems. Lucio Sansone is the author/co-author of more than 80 research papers in the field of Information Systems and he has been active in the field of distributed object systems, information retrieval, multimedia information systems and general techniques for the management of multimedia data, image analysis and description, object-oriented models for image processing. His current research interests lie in image and video databases.