Semantic Summarization of Web Documents

A. d'Acierno
ISA-CNR
Via Roma, 64
83100 Avellino, Italy
dacierno.a@isa.cnr.it

V. Moscato, F. Persia, A. Picariello
DIS - University of Naples
Via Claudio, 21
80125 Napoli, Italy
{vmoscato, fabio.persia, picus}@unina.it

A. Penta
DICAS - Politecnico di Torino
Castello del Valentino - Viale Mattioli, 39
10125 Torino, Italy
penta.antonio@gmail.com

***Abstract*—Documents' summarization techniques automatically extract relevant information from different sources with respect to a list of topics: they can be profitably used by a variety of applications and in particular for automatic indexing and categorization in order to facilitate the production and delivery of new multimedia contents. In this paper we propose a novel approach for summarizing documents retrieved from the Internet: we propose to capture the semantic nature of a document, expressed in natural language, in order to retrieve a number of RDF triplets and to clusterize these ones aggregating similar information. An overview of the system and some preliminary results are described.**

I. Introduction

It is commonly agreed both from scientific community and from common people using the Internet, that the process of accumulating and distributing information has a snowballing nature: in this continuous information generation process, about 80% of all reasonably accessible information in the web is contained in *natural-language* texts, sometimes strictly related to multimedia data and although sometimes the information is tremendously repeated.

This is surely one of the most important reasons for searching suitable and efficient *summarization* techniques: summarization, especially text summarization from web sources, may be considered as the process of "distilling" the most important information from a variety of logically related sources, as the one returned from classic search engines, in order to produce a short, concise and grammatically meaningful version of information spread out in pages and pages of texts.

Let us consider the work of a news reporter, that some times needs to read a certain number of articles for retrieving important information about a certain cool topic: this author needs - of course - to really understand and separate the most important ideas from the sometimes repeated info in the texts: it surely takes time and effort, and a great improvement could come if, instead of reading thousands and thousand of words, the news reporter reads a short article of few hundreds of words, so taking at least about ten or fifteen minutes! This could improve productivity as it also speeds up the surfing process and, instead of reading useless information, one could focus on summaries of web pages, of course in case they are precise and accurate.

From the point of view of modern information retrieval system, the use of summarization methods makes it possible to enhance both the accuracy and the relevance of retrieval: in addition, it provides useful possibilities of tracking the event and shift of a topic.

This kind of data reduction has become of great advantages for a variety of applications. The most important are especially related to several fields where information is a vital resource: in the biomedical domain, for example, physicians must continuously find clinical trial study information to incorporate into their patient treatment efforts and, the big problem is sured due to the high-volume of publications.

In all the previous examples, the ability to summarize a certain text document may be performed by a wide range of techniques, that takes into account language models (based on Natural Language Processing – *NLP* – theories) of different degrees of sophistication, performing, in other words, a "one size fits all" approach.

In this paper, we avoid the use of complex NLP techniques, and propose a novel approach for text summarization based on semantic extraction and description of documents. With the development of the Semantic Web, in fact, a large amount of native Resource Description Framework (RDF) documents are published on the Web and, for what concerns old pages, several well established techniques are able to transform a text document into a RDF triplet model.

In our approach, we propose to capture the semantic nature of a given document, expressed in Natural Language, in order to retrieve a number of RDF triplets and to clusterize these ones thus aggregating similar information. Our vision is that the more similar the RDF triplets are, the more the information is useless repeated: a summary may be so constructed using a sequence of RDF clusters, i.e. a sequence of representative triplet centroids.

In order to better understand our idea, let us consider the following running example, extracted from a news reported on an italian on line site after the crash of a piper in Milan.

"It was the second time since the Sept 11 terror attacks on New York and Washington that a plane has struck a high-rise building. Police and ambulances are at the scene. Many people were on the streets as they left work for the evening at the time of the crash. Ambulances streamed into the area and pedestrians peered upward at the sky. The clock fell to the floor. In Rome, a spokesman for the senate president, Marcello Pera, said the interior minister had informed him that the crash didn't appear to be a terror attack. – The president just moments ago was informed about the incident in Milan –

he said at his afternoon press briefing".

A RDF extractor, can easily find the following list, in the form: ⟨ subject, verb, object ⟩, the verb being reported in infinitive form:

⟨terror, attack, New York⟩, ⟨terror, attack, Washington⟩, ⟨plane, strike, building⟩, ⟨ambulance, be, scene⟩, ⟨police, be, scene⟩, ⟨people, be, street⟩, ⟨ambulance, stream, area⟩, ⟨pedestrian, peer, sky⟩, ⟨clock, fell, floor⟩, ⟨spokesman, say, minister⟩, ⟨minister, inform, crash⟩, ⟨president, inform, incident⟩.

A clustering algorithm may create the clusters reported in table I, where the retrieved centroids are underlined.

Cluster 1	⟨police, be, scene⟩ ⟨ambulance, be, scene⟩ ⟨people, be, street⟩ ⟨pedestrian peer sky⟩
Cluster 2	⟨minister, inform, crash⟩ ⟨president, inform, incident⟩ ⟨spokesman, say, minister⟩
Cluster 3	⟨clock, fell, floor⟩ ⟨plane, strike, building⟩ ⟨ambulance, stream, area⟩ ⟨terror, attack, Washington⟩ ⟨terror, attack, New York⟩

TABLE I
THE RESULTS OF THE CLUSTERING ALGORITHM ON A SET OF RDF TRIPLETS (CENTROIDS ARE UNDERLINED).

A summary is thus extracted considering the language sentences associated to each centroid: *Police and ambulances are at the scene. It was the second time since the Sept 11 terror attacks on New York and Washington that a plane has struck a high-rise building. In Rome, a spokesman for the senate president, Marcello Pera, said the interior minister had informed him that the crash didn't appear to be a terror attack.*

In the rest of the paper, we will describe the several steps for automatic summarization in order to obtain a similar summary from different and heterogeneous sources.

The paper is organized as follows. Section II describes the related work in the field of text summarization. Section III and section IV describe the proposed summarization process and the relative system; a case study is also shown and commented together with several experimental results in section V while conclusions and future works are discussed in section VI.

II. RELATED WORK

Summarization, i. e. the process of producing a shortened version of a document or of a set of documents, has been studied for a long time in the scientific community and automatic document summarization has been actively researched since the original work by Luhn [1] in 1958 . The summarization process usually produces an *extract*, when the sentences are preserved in their original form, or an *abstract*, when (at least part of) the content is generated using those sentences that are not present in the document(s). A further characterization is given by *indicative* summaries, that suggest the readers about the content of some documents, or by *informative* summaries, that are meant to substitute the original documents. Summaries can be also categorized into *query-based* (parts of the text are focused towards an information need expressed by a user using a query), *topic-* or *user-focused* (which contain only information on topics of interest for a well defined user), or *generic*.

In more recent years, a number of techniques have been proposed, using well extablished techniques in different fields, such as Pattern Recognition, Natural Language Processing and Semantic Analysis (see Semantic Vector Space models and Principal Component Analysis and Singular Value Decomposition [2] or *concept chains* [3], for examples).

Varadarajan et al. ([4], [5]) propose a method to add structure, in the shape of a graph, to text documents in order to allow effective *query specific* summarization. That is, they view a document as a set of interconnected text fragments (*passages*) and focus on keyword queries. The technique has the following key steps: first, at the preprocessing stage, they add a structure to each document, which can then be viewed as a labeled, weighted graph, called the *document graph*. Then, at query time, given a set of keywords, they perform keyword proximity search on the document graphs to discover how the keywords are associated in the document graphs. For each document, the summary is the minimum spanning tree on the corresponding document graph that contains all the keywords (or equivalent, based on a thesaurus).

Nowadays, due to the exponential growth of the Web, summarization can be profitably used to present HTML documents in condensed form and have their content in an handy shape; traditional techniques, on the other hand, may not capture the true meaning of a Web page for several reasons. First, one logical document may be presented on several pages or even on several sites linked each other so that page's HTML structure analysis is required. Then, Web documents contains media different from straight text, very important for human users but very difficult to summarize automatically; moreover, they often contain lots of unnecessary data and extracting relevant information is a key task that typically depends on user's needs. It is also to be noticed that, compared with a traditional document, a Web page is less constrained in its content, so that different topics are likely to be contained in a single page; another problem concern the fact that, given the intrinsic nature of the Web, HTML documents cannot be considered equally important.

In [6] a given number of HTML pages is collected from a given Web site via a breadth-first search starting at the home page and plain text is extracted and partitioned into paragraphs. Short paragraphs are filtered out, while long paragraphs are classified (using feature based classifiers) into narrative or non-narrative. Key-phrases are extracted from narrative text, anchor text and special text (e.g. italic text) using classifiers that learn the significance of each category of key-phrases. Finally, a summary is generated, scanning narrative paragraphs to extract the most significant sentences, as the ones containing a high density of key-phrases.

SWEeT ([7], Summarizer of WEb Topics) uses a user query

that is first passed to a search engine that answers with a set of relevant documents whose contents, together with some additional meta-information, are extracted and passed on to the summarizer. The summarizer extracts the most important sentences using a term-based strategy: finds the most important information in the document(s) by identifying its main terms, and then extracts from the document(s) the most important information (i.e., sentences) about these terms. A latent semantic analysis is then used to reduce the dimensionality of the term space and similar terms and sentences are clustered into *topics*. Sentences with the most important topics are then selected for the summary.

Finally, *Importance* (that evaluates how the summary takes into account the user keywords), *Consistence* (a criterion that gives more importance to those blocks of consecutive sentences coming from the same sources) and *Recurrence* (to measure how semantically similar are the concepts presented in the sentences) have been used [8] in a summarizer that applies several optimization criterion.

Differently from the previous works, the method presented in this paper proposes a novel methodology based on the semantic extraction of information and on the semantic clustering of similar sentences.

III. The Summarization Process

In our vision, the summarization process determines,with respect to a set of textual sentences, a *summary* that synthesizes the related semantic content, applying a certain *summarization function*.

We think that a summary may be usefully derived considering a RDF representation of the source texts instead of the sources themselves. In particular, we represents each sentence in a source text, or *summarizable sentence*, as a triple $\sigma = \langle s, \tau, \omega \rangle$, s being the original sentence, τ the relative *RDF-triple* describing the semantic content of s in terms of *subject*, *predicate* and *object* and ω an identifier related to the page source containing s. In this model, τ can be obtained using different *information extraction and processing* algorithms.

In addition, a large number of sophisticate not-exact semantic distances between RDF triplets have been proposed in the literature, ranging from distance-oriented measures computed on apposite dictionaries, to metrics based on models of distributional similarity learned from large text collections (e.g. Leacock & Chodorow, Lesk, Wu & Palmer [9], Resnik, Dekang-Lin [10] and Jiang & Conrath [6]).

In other words, each sentence is mapped into a *metric space* whose dimensions are the related subject, predicate and object, and a relative distance metric: summarization problem can be thus formalized as the problem of determining in this metric space a set of triples, or a *summary*, that better represents the semantics of the document.

The proposed summarization process is schematically presented in the following.

1) Textual sentences are extracted from web pages by parsing HTML code. In particular:
   a) useful text is detected by analyzing HTML tags;
   b) the related sentences are extracted and anaphor crossing references are solved.
2) The *Named Entities* of each sentences are recognized.
3) Subject, predicate and object of each extracted sentence are detected by analyzing the related *Parse Tree*. In particular:
   a) the parse tree is generated (figure 1 shows the parse tree related to the sentence "A small plat has hit a skyscraper in central Milan");
   b) apposite heuristic *search patterns*, based on relationships among nodes and on their kind, are applied on the parse tree in order to discover subject, predicate and object.
4) The discovered subject, predicate and object are extracted from each sentence and represented in a subject-predicate-object space by a RDF format.
5) A matrix containing the semantic distance for each couple of summarizable sentences is computed.
6) A clustering algorithm is applied as summarization function in the subject-predicate-object space.
7) The sentences related to the representative of each clusters are organized in a summary.

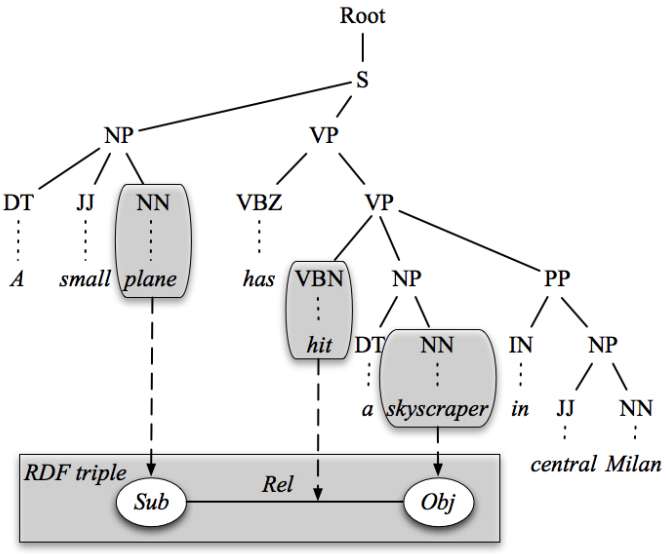


Fig. 1. An example of extraction of RDF triple from parse tree generation.

IV. System Implementation

In our vision, a web summarization system is very similar to a classic search engine: a user, by means of several keywords, activates a *search engine* that stores in a repository all web pages satisfying the searching criteria. A *NLP* module processes the stored web pages and extracts, for each one, the related textual sentences. More in details, in a first step useful text of web pages is extracted by parsing HTML code by means of *Sentences Extraction* component that we implemented using *Jericho HTML Parser*. In a second stage,

appropriate NLP algorithms perform the text splitting into sentences and the recognition of related named entities using the *Named Entity Recognition* and the *Anaphora Resolution* components of *Stanford NLP Libraries*. A *Triples Management* module extracts the related semantic content from original sentences in terms of RDF triples. In particular, the *Triples Extraction* component extracts from each sentence the related RDF representation by analyzing the Parse Tree: the tree generation is obtained by using an enhanced version of the *Stanford Parser*, while the tree inspection is realized by the *Stanford Tregex Libraries*. A *Triples Clustering* component performs clustering of the RDF triples: in our implementation, we used a clustering algorithm based on the well known *K-Medoid* approach.

To accomplish the clustering task, we need a *Semantic Distance Computation* module that computes, by using the *Wordnet* lexical database and *Semantic Distance* components, the semantic distances for each couple of summarizable sentences. In our implementation, we adopted Wu & Palmer [9] and Dekang-Lin distances [10]. Finally, a *Summary Management* module has the task of generating the best summary related to the search keywords. By means of the *Summary Sentences Extraction* component, sentences related to the representative of each clusters are organized in a summary. The order of the sentences is based on the average distance between representatives and the other elements of a clusters: the most important sentences are those related to the cluster having the lower values of distance.

V. Case Study and Experimental Results

The case study presented in this section concerns to a possible scenario in which a news reporter needs to have some information delivered to his PDA or cell-phone for writing an article outlining "Pirelli tower plane crash", when a little airplane crashed to the Pirelli skyscraper in Milan. He does not want to waste time analyzing a great amount of information sources and for such a reason a summarization system could helpfully produce a short summary of the gathered retrieved information, and in addition, such a summary could be easily managed and accessed using light mobile devices. By means of our summarization system, the reporter can directly select some web sources (e.g. news sites) describing the air accident and request to obtain several short summaries of different lengths about the chosen topic, thus avoiding the boring drawback of searching useful information site by site and page by page.

In table II, we show an example of different size summaries generated by the system in correspondence of a subset of textual sentences coming from the selected web sources. For simplicity and readability sake, we selected two web sources, each related to a certain news web portal; it is possible to observe as the sentences reflecting similar semantics are clustered in the RDF space and then represented by the related medoid.

In order to evaluate our system, we have used 4 different data corpora used in the *text-summarization* field. The first (herein called *Milan*) contains 9 papers about the air accident happened in Milan the 18th April 2002, when a little airplane crashed to the Pirelli skyscraper; there are also 3 ideal summaries generated by humans (15%, 25 % and 35% of the original document length). The second data corpus (*Moscow*) is composed of 4 papers about a bomb explosion in a Moscow building, which caused many dead and injured; three human-generated summaries (15%, 25 % and 35% of the original document length) are also available. A document set containing 3 papers about Al-Qaida with a summary whose length is 38% of the original document length is the third data set we used in our experiments (*Al-Qaida*) while the fourth data corpus (*Turkey*) is a document about the problems pertinent to the Turkey attachment to European Community with and available summary with a number of sentences equal to the 33% of the original document length.

The implemented system has been tested both using the *Dekang Ling* metric with the weights distribution 0.33-0.33-0.33 (*Lin33*) and the *Wu&Palmer* metric with the weight distribution 0.40-0.30-0.30 (*Wup40*)[1]. In addition, we have compared our experimental results with three commercial systems (Copernic Summarizer [11], QuickJist [12] and Search Subject Summarizer [13]) and with two free available applications (Online Summarization [14] and Great Summary [15]), that are based on different strategies and algorithms rather then on RDF clusterings: in this way, we want to have an idea of how much robust and efficient is our algorithm in capturing the main concepts of a set of documents.

The summarizers have been first tested on long summaries (i.e. longer than 30% of the original document); obtained results on the available data sets are oulined in table V.a where systems are organized as the average performance (AVG) decreases. Our systems outperforms the freely available systems and we obtain also summaries that are better than the *Copernic Summarizer* summary. *QuickJist* and *Search Subject Summarizer* show better summaries, even if we obtain very similar results except for the *Moscow* data corpus. For summaries whose length is between the 20% and the 30% of the original document's length (see table V.b) we again outperform the freely available systems and obtain better results than *Copernic Summarizer*. In this case the *Dekang Ling* metric works slightly better than the *Wu&Palmer* one. When we consider short summaries (i.e. with a length less than the 20% of the original length) we obtain the results depicted in table V.c; here, *Wup40* on average outperforms each system with the exception of *QuickJist*. The obtained co-occurrence similarities show, not surprisingly, that it is very difficult to generate good short summaries.

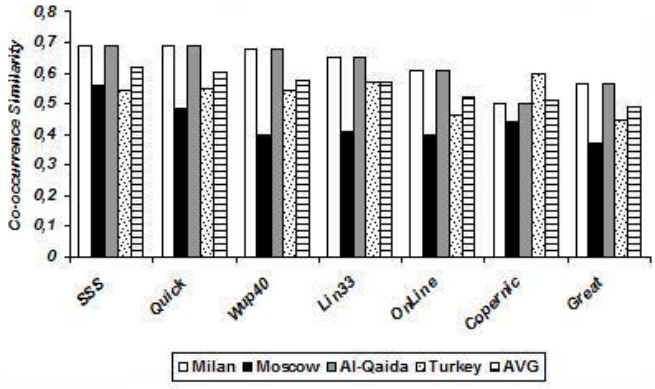
VI. Conclusions

In this paper we described a novel approach for summarizing web documents retrieved from the Internet based on semantic extraction and RDF description of documents. In particular, we proposed a system able to build summaries

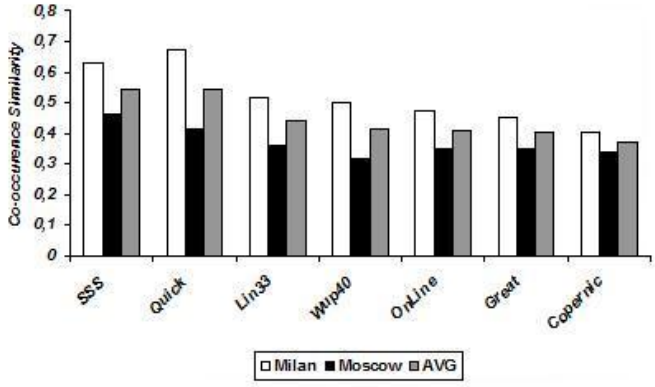
[1]These parameters have been experimentally tuned.

First web source
A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN. The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina. The building houses government offices and is next to the city's central train station. Several storeys of the building were engulfed in fire, she said. Italian TV showed a hole in the side of the Pirelli building with smoke pouring from the opening. RAI state TV reported that the plane had apparently radioed an SOS because of engine trouble. Earlier though, in Rome, the senate's president, Marcello Pera, said it "very probably" appeared to be a terrorist attack. Police and ambulances are at the scene. Many people were on the streets as they left work for the evening at the time of the crash. Police were trying to keep people away, and many ambulances were on the scene. There is no word yet on casualties. TV pictures from the scene evoked horrific memories of the September 11 attacks on the World Trade Center in New York and the collapse of the building's twin towers. "I heard a strange bang so I went to the window and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them," said Gianluca Liberto, an engineer who was working in the area told Reuters. The building is known as the Pirelli skyscraper but the Italian tyre and cable company does not operate out of the building. It is one of the symbols of Italy's financial capital and is one of the world's tallest concrete buildings, designed between 1955 and 1960. U.N. envoy horror at Jenin camp U.S. bombing kills Canadians Chinese missiles concern U.S. A small plane crashed into the tallest building in downtown Milan Thursday evening, causing smoke to pour out of the top floors of the skyscraper. Three people were reported to have been killed, and rescue workers said dozens of people had been taken to a nearby hospital. Officials said the crash appeared to have been an accident and that only the pilot was aboard. The weather was clear at the time of the crash. Police officer Celerissimo De Simone said the pilot of the Piper aircraft had sent out a distress call at 5:50 p.m. just before the crash near Milan's main train station. RAI state TV reported that the pilot said the SOS was because of engine trouble. The plane had taken off from Locarno, Switzerland, and was heading to Milan's Linate airport, De Simone said. Earlier, in Rome, the senate's president, Marcello Pera, said it "very probably" appeared to be a terrorist attack. But Pera's spokesman later said he had spoken with the Interior Minister and the crash didn't appear to be any kind of an attack.
Second web source
A small tourist plane hit a skyscraper in central Milan this afternoon, setting the top floors of the 30-storey building on fire. A spokesman for Italy's air safety agency said the crash was an accident and not a terror attack. It was clearly an air accident, Adalberto Pellegrino of the National Air Safety Agency confirmed. He said there appeared to be no foul play. Italian transport officials said the pilot, named as 75-year-old Luigi Fasulo, reported technical problems before hitting the skyscraper. The crash, which happened at 5.45pm local time(3.45pm GMT) killed at least three people and several dozen were injured a senior local government official confirmed. "We have news of three dead and dozens and dozens of injured," said Pier Gianni Prosperini, the deputy head of Milan's regional government. Police said one woman leapt to her death as smoke billowed from the scarred top of the skyscraper, which houses the Lomabardy regional government offices. Civil servant Maurizio Sala was on the 20th floor when two explosions shook his building. "We all rushed to the window and we suddenly realised it was something similar to the World Trade towers because thousands of pieces of paper were flying through the air. It was the same image," he said. A regional official said that by a stroke of luck the top five floors of the Milan skyscraper were virtually empty when the plane struck at around because they were being renovated. It's also believed around 1,000 of the 1,300 people who work in the skyscraper had already left to go home for the evening. Many of the injured were hit by falling debris outside the building. Eyewitnesses reported hearing a loud explosion from the office block which sits next to the city's central train station. "I heard a strange bang so I went to the windows and outside I saw the windows of the Pirelli building blown out and then I saw smoke coming from them," said Gianluca Liberto, an engineer who was working in the area. Police have identified the pilot of the four-seater Rockwell Commander plane as 75-year-old Luigi Fasulo. Fasulo is believed to have had some 30 years flying experience and took off from the southern Swiss town of Locarno bound for Linate airport on the outskirts of Milan, transport officials said. The building that was hit is known as the Pirelli skyscraper but the Italian tyre and cable company does not operate from it. It is one of the symbols of Milan, Italy's financial capital, and stands 127 metres (400 feet) tall. The city's stock exchange suspended share trading after the incident. First reports of the crash inspired fears of a terrorist attack. Since the September 11 hijacked airliner attacks on New York's World Trade Center and in Washington, Italy has been at the forefront of the U.S.-led war on terrorism in Europe. In October, U.S. officials said they believed Milan's Islamic Cultural Institute was al Qaeda's main European base. Muslim leaders in Italy have denied that charge. Italy has arrested around 30 people on suspicion of links to extremist Islamic groups since September 11 and has frozen around 345 million euros (300 million) of suspected assets. In February a terrorist plot was uncovered against the US Embassy in Rome and at Easter, the Embassy issued an official warning that Italian cities, including Milan could be the subject of a terrorist attack over the Easter weekend.
A summary of 20 sentences
A small tourist plane hit a skyscraper in central Milan this afternoon, setting the top floors of the 30-storey building on fire. The crash, which happened at 5 45pm local killed at least three people and several dozen were injured a senior local government official confirmed. Police said one woman leapt to her death as smoke billowed from the scarred top of the skyscraper, which houses the Lomabardy regional government offices. "We all rushed to the window and we suddenly realised it was something similar to the World Trade towers thousands of pieces of paper were flying through the air. A regional official said that by a stroke of luck the top five floors of the Milan skyscraper were virtually empty when the plane struck at around they were being renovated. Many of the injured were hit by falling debris outside the building. The building that was hit is known as the Pirelli skyscraper but the Italian tyre and cable company does not operate from it. First reports of the crash inspired fears of a terrorist attack. Since the September 11 hijacked airliner attacks on New York's World Trade Center and in Washington, Italy has been at the forefront of the U. S. -led war on terrorism in Europe. Muslim leaders in Italy have denied that charge. Italy has arrested around 30 people on suspicion of links to extremist Islamic groups since September 11 and has frozen around 345 million euros of suspected assets. In February a terrorist plot was uncovered against the US Embassy in Rome and at Easter, the Embassy issued an official warning that Italian cities, including Milan could be the subject of a terrorist attack over the Easter weekend. The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p. m. on Thursday, said journalist Desideria Cavina. The building houses government offices and is next to the city's central train station. Italian TV showed a hole in the side of the Pirelli building with smoke pouring from the opening. Earlier though, in Rome, the senate's president, Marcello Pera, said it "very probably" appeared to be a terrorist attack. The building is known as the Pirelli skyscraper but the Italian tyre and cable company does not operate out of the building. Three people were reported to have been killed, and rescue workers said dozens of people had been taken to a nearby hospital. Police officer Celerissimo De Simone said the pilot of the Piper aircraft had sent out a distress call at 5:50 p. m. just before the crash near Milan's main train station. Earlier, in Rome, the senate's president, Marcello Pera, said it "very probably" appeared to be a terrorist attack.
A summary of 8 sentences
A small tourist plane hit a skyscraper in central Milan this afternoon, setting the top floors of the 30-storey building on fire. Civil servant Maurizio Sala was on the 20th floor when two explosions shook his building. A regional official said that by a stroke of luck the top five floors of the Milan skyscraper were virtually empty when the plane struck at around because they were being renovated. Fasulo is believed to have had some 30 years flying experience and took off from the southern Swiss town of Locarno bound for Linate airport on the outskirts of Milan, transport officials said. First reports of the crash inspired fears of a terrorist attack. Muslim leaders in Italy have denied that charge. Italian TV showed a hole in the side of the Pirelli building with smoke pouring from the opening.

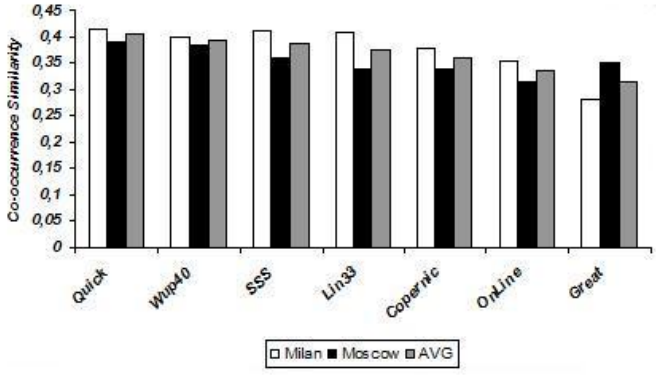
TABLE II
A CASE STUDY



(a)



(b)



(c)

TABLE III
EXPERIMENTAL RESULTS ON LONG SUMMARIES (A), ON MEDIUM SUMMARIES (B) AND ON SHORT SUMMARIES (C).

by using sequences of clusters samples in the RDF space. We tested our system using some well known data corpora publicly available and we obtained promising results both with respect to freely available and commercial applications. Further works will be devoted to improve our work into three main directions: i) extend the methodology to multimedia data; ii) design more detailed experiments and iii) compare the results with other similar summarization systems.

ACKNOWLEDGMENT

This work has been carried out partially under the financial support of the Programma Faro (Finanziamento per lAvvio di Ricerche Originali) in the framework of the LATINO Project: Un sistema innovativo di LocAlizzazione e Tracciamento di INdividui per l'Ottimizzazione di percorsi in ambienti indoor.

REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, 1958.

[2] O. Vikas, A. K. Meshram, G. Meena, and A. Gupta, "Multiple document summarization using principal component analysis incorporating semantic vector space model," *Computational Linguistics and Chinese Language Processing*, vol. 13, no. 2, pp. 141–156, 2008.

[3] L. Yu, J. Ma, F. Ren, and S. Kuroiwa, "Automatic text summarization based on lexical chains and structural features," *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, ACIS International Conference on*, vol. 2, pp. 574–578, 2007.

[4] R. Varadarajan and V. Hristidis, "Structure-based query-specific document summarization," in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2005, pp. 231–232.

[5] ——, "A system for query-specific document summarization," in *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2006, pp. 622–631.

[6] Y. Zhang, N. Zincir-Heywood, and E. Milios, "World wide web site summarization," *Web Intelli. and Agent Sys.*, vol. 2, no. 1, pp. 39–53, 2004.

[7] J. Steinberger, K. Jezek, and M. Sloup, "Web topic summarization," in *Proceedings of the 12th International Conference on Electronic Publishing*, June 2008, pp. 322–334.

[8] A. d'Acierno, V. Moscato, and A. Picariello, "Building summaries from web information sources," in *WIAMIS '09: Proceedings of the 10th Workshop on Image Analysis for Multimedia Interactive Services*. IEEE Computer Society, 2009, pp. 57–60.

[9] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, 1994, pp. 133 –138. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.1869

[10] D. Lin, "An information-theoretic definition of similarity," in *ICML*, J. W. Shavlik, Ed. Morgan Kaufmann, 1998, pp. 296–304.

[11] "http://www.copernic.com/en/products/summarizer/index.html."

[12] "http://www.fileguru.com/quickjist-summarizer/info."

[13] "http://www.kryltech.com/summarizer.htm."

[14] "http://www.tools4noobs.com/summarize/."

[15] "http://www.greatsummary.com/."