Chapter 4

A Semantic-Based Framework for Summarization and Page Segmentation in Web Mining

Alessio Leoncini, Fabio Sangiacomo,
Paolo Gastaldo and Rodolfo Zunino

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51178

1. Introduction

The World Wide Web has become a fundamental resource of information for an increasing number of activities, and a huge information flow is exchanged today through the Internet for the widest range of purposes. Although large-bandwidth communications yield fast access to virtually any kind of contents by both human users and machines, the unstructured nature of most available information may pose a crucial issue. In principle, humans can best extract relevant information from posted documents and texts; on the other hand, the overwhelming amount of raw data to be processed call for computer-supported approaches. Thus, in recent years, *Web mining* research tackled this issue by applying data mining techniques to Web resources [1].

This chapter deals with the predominant portion of the web-based information, i.e., documents embedding natural-language text. The huge amount of textual digital data [2, 3] and the dynamicity of natural language actually can make it difficult for an Internet user (either human or automated) to extract the desired information effectively: thus people every day face the problem of information overloading [4], whereas search engines often return too many results or biased/inadequate entries [5]. This in turn proves that: 1) treating web-based textual data effectively is a challenging task, and 2) further improvements are needed in the area of Web mining. In other words, algorithms are required to speed up human browsing or to support the actual crawling process [4]. Application areas that can benefit from the use of these algorithms include marketing, CV retrieval, laws and regulations exploration, competitive intelligence [6], web reputation, business intelligence [7], news articles search [1], topic tracking [8], and innovative technologies search. Focused crawlers represent another potential, crucial area of application of these technologies in the security domain [7, 9].

The research described in this chapter tackles two challenging problems in Web mining techniques for extracting relevant information. The first problem concerns the acquisition of useful knowledge from textual data; this is a central issue for *Web content mining* research, which mostly approached this task by exploiting text-mining technologies [1]. The second problem relates to the fact that a web page often proposes a considerable amount of information that can be regarded as 'noise' with respect to the truly informative sections for the purposes at hand [10]. According to [10], uninformative web page contents can be divided into navigation units, decoration items, and user interaction parts. On one hand, these elements drain the user's attention, who has to spend his/her time to collect truly informative portions; on the other hand, they can affect the performances of algorithms that should extract the informative content of a web page [10]. This problem is partially addressed by the research area of *semantic Web*, which aims to enrich web pages with semantic information accessible from humans and machines [5]. Thus *semantic Web mining* aims to combine the outcomes of semantic Web [11] and Web mining to attain more powerful tools that can reliably address the two problems described above [5].

The approach adopted in this work, however, does not rely on semantic information already embedded into the Web resources, and the semantic characterization of words and sentences plays a crucial role to reach two outcomes:

- to work out from a Web resource a concise summary, which outlines the relevant topics addressed by the textual data, thus discarding uninformative, irrelevant contents;
- to generate a web page segmentation that points out the relevant text parts of the resource.

Semantic characterization is obtained by applying semantic networks to the considered Web resource. As a result, natural language text maps into an abstract representation, that eventually supports the identification of the topics addressed in the Web resource itself. A heuristic algorithm attains the latter task by using the abstract representation to work out the relevant segments of text in the original document. Page segmentation is then obtained by properly exploiting the information obtained on the relevant topics and the topics covered by the different sections of the Web page.

The novelty contribution of this work lies in a framework that can tackle two tasks at the same time: text summarization and page segmentation. This result is obtained by applying an approach that extracts semantic information from the Web resource and does not rely on external information that may not be available. Combining effective page segmentation with text summarization can eventually support advanced web content mining systems that address the discovery of patterns, the tracking of selected topics and the efficient resource finding.

Experimental results involved the well-know DUC 2002 dataset [12]. Such dataset has been used to evaluate the ability of the proposed framework to consistently identify the topics addressed by a document and eventually generate the corresponding summary. The ROUGE tool [13] has been used to measure the performance of the summarization algorithm exploited by the present framework. Numerical results proved that the research described in this chapter compares positively with state-of-the-art approaches published in the literature.

The rest of the chapter is organized as follows. Section 2 gives an overview of the state of the art in the different research areas involved. Section 3 introduces the overall approach proposed in this research, while Section 4 discusses the actual implementation of the framework. Section 5 presents the experimental results. Some concluding remarks are made in Section 6.

2. Related work

The current research proposes a web mining algorithm that exploits knowledge-based semantic information to integrate text-summarization and web page-segmentation technologies, thus improving the overall approach effectiveness. The following sections overview the state of the art in the different research areas involved: web content mining, text summarization, and web page segmentation. The Section also highlights the points of novelty introduced by the present research with respect to previous works.

2.1. Web content mining

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services; the applicative areas include resource finding, information selection, generalization and data analysis [14]. Incidentally, machine-learning methods usually address the last two tasks. Web mining includes three main sub-areas: web content mining, web structure mining, and web usage mining [15]. The former area covers the analysis of the contents of web resources, which in general comprise different data sources: texts, images, videos and audio; metadata and hyperlinks are often classified as text content. It has been proved that unstructured text represents the prevailing part of web resources [14, 16] this in turn motivates the large use of text mining technologies.

A wide variety of works in the literature focused on text mining for web content mining [17]. Some web content mining techniques for web search, topic extraction and web opinion mining were explored in [18]. In [19], Liu et al. showed that web content mining could address applicative areas such as sentiment classification, analysis and summarization of consumer reviews, template detection and page segmentation. In [20], web content mining tackled business applications by developing a framework for competitive intelligence. In [21], an advanced search engine supported web-content categorization based on word-level summarization techniques. A web-page analyzer for detecting undesired advertisement was presented in [22]. The work described in [23] proposed a web-page recommendation system, where learning methods and collaborative filtering techniques cooperated to produce a web filter for efficient user navigation.

The approach presented in this research differs from those related works in two main aspects: first, it exploits semantic-based techniques to select and rank single sentences extracted from text; secondly, it combines summarization with web page segmentation. The proposed approach does not belong to the semantic web mining area, which refers to methodologies that address the development of specific ontologies that enrich original web page contents in a structured format [11, 24]. To the best of the authors' knowledge, the literature provides only two works that used semantic information for web content mining. The research described in [25] addressed personalized multimedia management systems, and used semantic, ontology-based contextual information to attain a personalized behavior in content access and retrieval. An investigation of semantic-based feature extraction for web mining is proposed in [26], where the WordNet [27] semantic network supported a novel metrics for semantic similarity.

2.2. Text summarization

A summary is a text produced by one or more other texts, expressing important information of original texts, and no longer than half of the original texts [28]. Actually, text summarization techniques aim to minimize the reading effort by maximizing the information density that is prompted to the reader [29]. Summarization techniques can be categorized into two approaches: in extractive methods, summaries stem from the verbatim extraction of words or sentences, whereas abstractive methods create original summaries by using natural language generators [30].

The works of Das et al. [30] and Gupta et al. [31] provided extensive surveys on extractive summarization techniques. Several methods relied on word frequency analysis, cue words extraction, or selection of sentences according to their position in the text [32]. More recent works used tf-idf metrics (term frequency - inverse document frequency) [33], graphs analysis, latent semantic analysis [34], machine learning techniques [35], and fuzzy systems [36, 37]. Other approaches exploited semantic processing: [38] adopted lexicon analysis, whereas concepts extraction supported the research presented in [39]. Abstractive summarization was addressed in [40], where the goal was to understand the main concepts of a document, and then to express those concepts in a natural-language form.

The present work actually relies on a hybrid extractive-abstractive approach. First, most informative sentences are selected by using co-occurrence of semantic domains [41], thus involving an extractive summarization. Then, abstractive information is produced by working out the most representative domains for every document.

2.3. Web page segmentation

Website pages are designed for visual interaction, and typically include a number of visual segments conveying heterogeneous contents. Web page segmentation aims to grasp the page structure and split contents according to visual segments. This is a challenging task that brings about a considerable number of issues. Different techniques were applied to web page segmentation in the past years: PageRank [42], graphs exploration [43], rules [10, 44, 45], heuristics [46, 47, 48, 49], text processing [50], image processing [51], machine learning [52, 53], and semantic processing [54].

Web page segmentation methods apply heuristic algorithms, and mainly rely on the Document Object Model (DOM) tree structure that is associated to a web resource. Therefore, segmentation algorithms may not operate properly when those ancillary features are not available or when they do not reflect the actual semantic structure of the web page. Conversely, the approach presented in this chapter only relies on the processing of the textual information that can be retrieved in the web resource.

3. A Framework for Text Summarization and Segmentation

The processing of textual data in a Web page yields two outcomes: a text summary, that identifies the most relevant topics addressed in the Web page, and the set of sentences that are most correlated with those topics. The latter indirectly supports the segmentation of the web page, as one can identify the substructures that deal with the relevant topics. Several advanced applications for Web mining can benefit from this approach: intelligent crawlers that explore links only related to most informative content, focused robots that follow specific content evolution, and web browsers with advertising filters or specific content- highlighting capabilities. This Section presents the overall approach, and introduces the various elements that compose the whole framework. Then, Section 4 will discuss the actual implementation of the framework used in this work.

3.1. Overall system description

The approach relies on a two-level abstraction of the original textual information that is extracted from the web page (Figure 1); semantic networks are the tools mainly exploited to accomplish this task. First, raw text is processed to work out *concepts*. Then, concepts are grouped into domains; here, a domain represents a list of related words describing a particular subject or area of interest. According to Gliozzo et al [55], domain information corresponds to a paradigmatic relationship, i.e., two words with meanings that are closely related (e.g., synonyms and hyponyms).

Semantic networks allow to characterize the content of a textual resource according to semantic domains, as opposed to a conventional bag of words. The ultimate objective is to exploit a coarse-grained level of sense distinctions, which in turn can lead to identify the topics actually addressed in the Web page. Toward that end, suitable algorithms must process the domain-based representation and recognize the relevant information in the possibly noisy environment of a Web page. Indeed, careful attention should be paid to the fact that many Web pages often address multiple, heterogeneous domains. Section 4 presents in detail the procedure implementation to identify specific domains in a Web page.

Text summarization is obtained after the identification of the set, Θ, of domains that characterize the informative content of the Web page. The summary is obtained by detecting in the original textual source the sentences that are mostly correlated to the domains included in Θ. To complete this task sentences are ranked according to the single terms they involve, since the proposed approach only sets links between terms and concepts (domains). The process can generate the eventual summary according to two criteria: the first criterion yields a summary that describes the overall content of the Web page, and therefore does not distinguish the various domains included in Θ; the second criterion prompts a multiplicity of summaries, one for each domain addressed in Θ.

That approach to text summarization supports an unsupervised procedure for page segmentation, too. Indeed, the described method can 1) identify within a Web page the sentences that are most related to the main topics addressed in the page itself, and 2) label each sentence with its specific topic. Thus text summarization can help assess the structure of the Web page, and the resulting information can be combined with that provided by specific structure-oriented tools (e.g., those used for tag analysis in html source code).

Figure 2 shows the two alternative strategies that can be included in the Web mining system. The first strategy uses the text summarization abilities to find relevant information in a Web page, and possibly to categorize the contents addressed. The second strategy targets a selective search, which is driven by a query prompted by the user. In the latter case, text summarization and the eventual segmentation allow the mining tool to identify the information that is relevant for the user in the considered Web page.

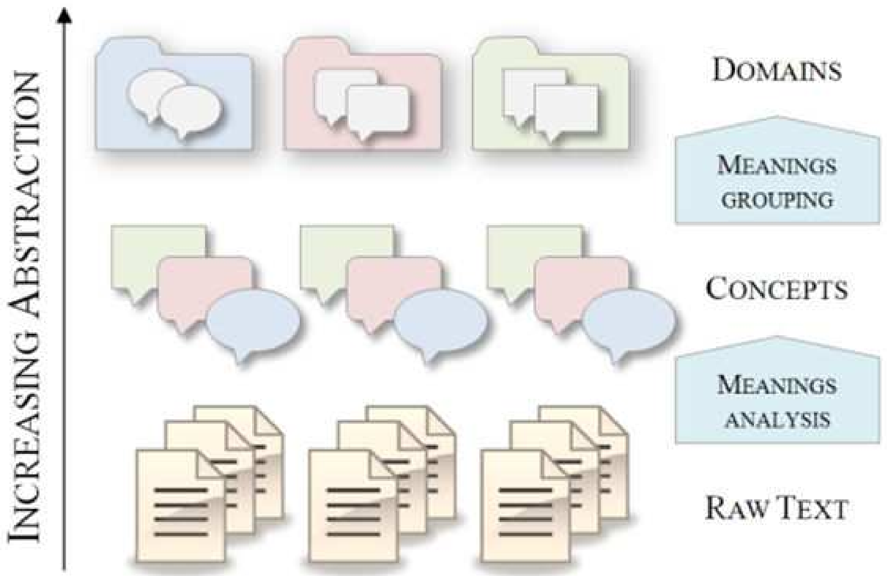


Figure 1. The two abstraction layers exploited to extract contents from textual data.

3.2. Overall system description

The overall framework can be schematized according to the following steps (Figure 3):

From the Web page to textual data:

a. get a Web page;

b. iextract textual data from the source code of the Web page.

Text preprocessing:

a. identify words and sentences terminators to split text into words (tokens) and sentences;

b. erase stop words;

c. lemmatization.

Abstraction:

a. first abstraction level: a semantic network is used to extract a set of concepts from every token; eventually, a list of concepts is obtained;

b. second abstraction level: the concepts are grouped in homogeneous sets (domains).

Content analysis:

a. strategy: automatic selection of domain

b. identify the informative contents addressed by processing the list of domains obtained after Step 3 (Abstraction);

c. strategy: user-driven domain

process the list of domains obtained after Step 3 (Abstraction) to search for the topics indicated by the user.

Outputs:

Summarization:

a. use the output of Step 4 (Content Analysis) to rank the sentences included in the textual source;

b. build a summary by using the most significant sentences according to the rank.

Page Segmentation:

a. use the sentences ranking to select the portions of the web page that deal with the main topics.

Step 4 (Content Analysis) and Step 5 (Outputs) can be supported by different approaches. Section 4 will discusses the approaches adopted in this research.

4. Implementation

The processing starts by feeding the system with the download of a web page. Raw text is extracted by applying the ‘libxml’ parsing library [56] to the html source code.

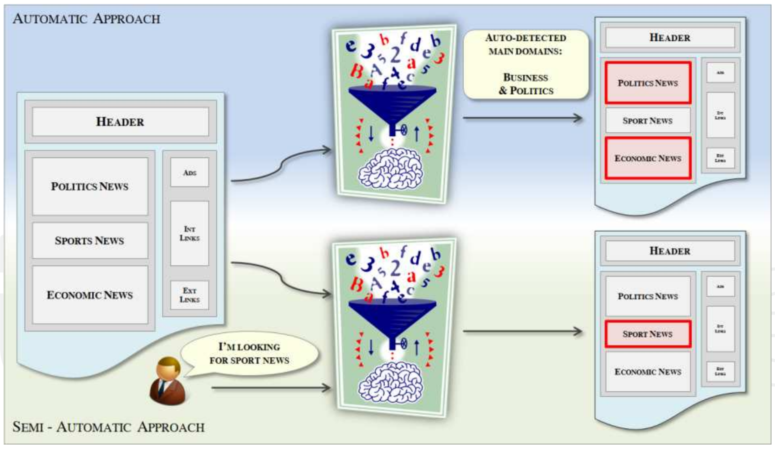


Figure 2. The proposed system can automatically detect the most relevant topics, or alternatively can select single text sections according to the user requests

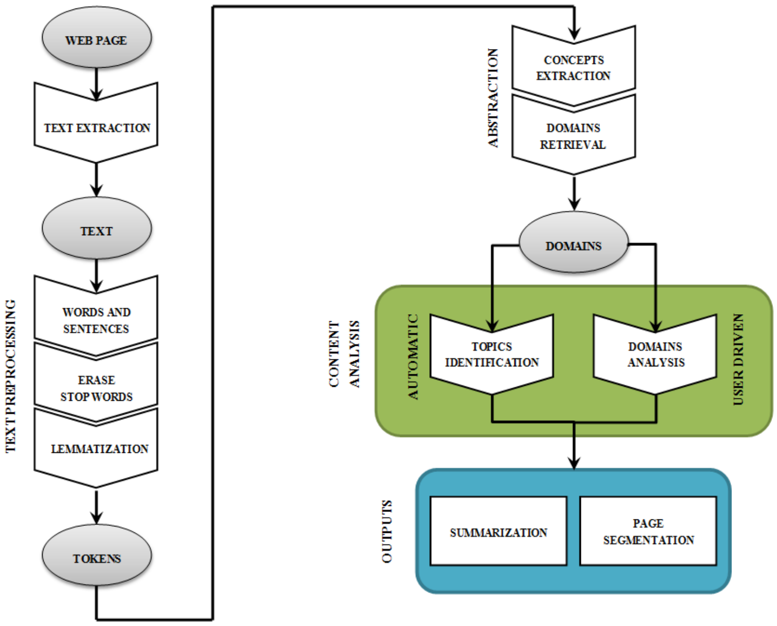


Figure 3. The data flow of the proposed framework

4.1. Text preprocessing

This phase receives as input the raw text and completes two tasks: 1) it identifies the beginning and the end of each sentence; 2) it extracts the tokens from each sentence, i.e., the terms that compose the sentence. Additional subtasks are in fact involved for optimal text processing: after parsing raw text into sentences and tokens, idiom is identified and stop-words are removed accordingly; this operation removes frequent and semantically non-selective expressions from text. Then, *lemmatization* simplifies the inflectional forms of a term (sometimes derivationally related forms) down to a common radix form (e.g., by simplifying plurals or verb persons). These subtasks are quite conventional in natural language processing systems [57], and aim to work out a set of representative tokens.

The process that extracts sentence and tokens from text is driven by a finite-state machine (FSM), which parses the characters in the text sequentially. The formalism requires the definition of the following quantities:

- state *STARTT*: a token begins;
- state *ENDT*: end of token achieved;
- state *STARTS*: a sentence begins (hence, also a token begins);
- state *ENDS*: end of sentence achieved (hence, end of token also achieved);
- set *tdelim*, which includes space, tab and newline codes, plus the following characters: “\’,/:;.!?[]{}()*^-~_=
- set *sdelim*, which includes common sentence delimiter characters, such as :;!?’”
- set *number*, which includes all the numbers;
- set *lower*, which includes all the lower case alphabet characters;
- set *upper*, which includes all the upper case alphabet characters;
- set character, which is obtained as the union of set *lower* and set *upper*;
- set *dot*, which only include the dot character.

A detailed description of the complete procedure implemented by the FSM is provided in Figure 4. Actually, Figure 4(a) refers to the core procedure, which includes the initial state STARTS; Figure 4(b) refers to the sub-procedure that starts when the state NUMBER is reached in the procedure of Figure 4(a); Figure 4(c) refers to the sub-procedure that starts when the state ALPHA is reached in the procedure of Figure 4(a). In all the schemes the elements with circular shape represent the links between the three procedures: the light-grey elements refer to links that transfer the control to a different procedure; the dark-grey elements refer to links that receive the control from a different procedure.

The process implemented by the FSM yields a list of tokens, a list of sentences and the position of each token within the associated sentence. Stop-word removal takes out those tokens that either are shorter than three characters or appear in a language-specific list of terms (conjunctions, articles, etc). This effectively shrinks the list of tokens. Finally, a lemmatization process reduces each token to its root term. Different algorithms can perform the lemmatization step, depending on the document language. WordNet morphing features [27] support best lemmatization in the English idiom, and has been adopted in this research.

In the following, the symbol Ω will define the list of tokens extracted after text preprocessing: $\Omega = \{t_i; i = 1,..,N_t\}$, where t_i is a token and N_t is the number of tokens.

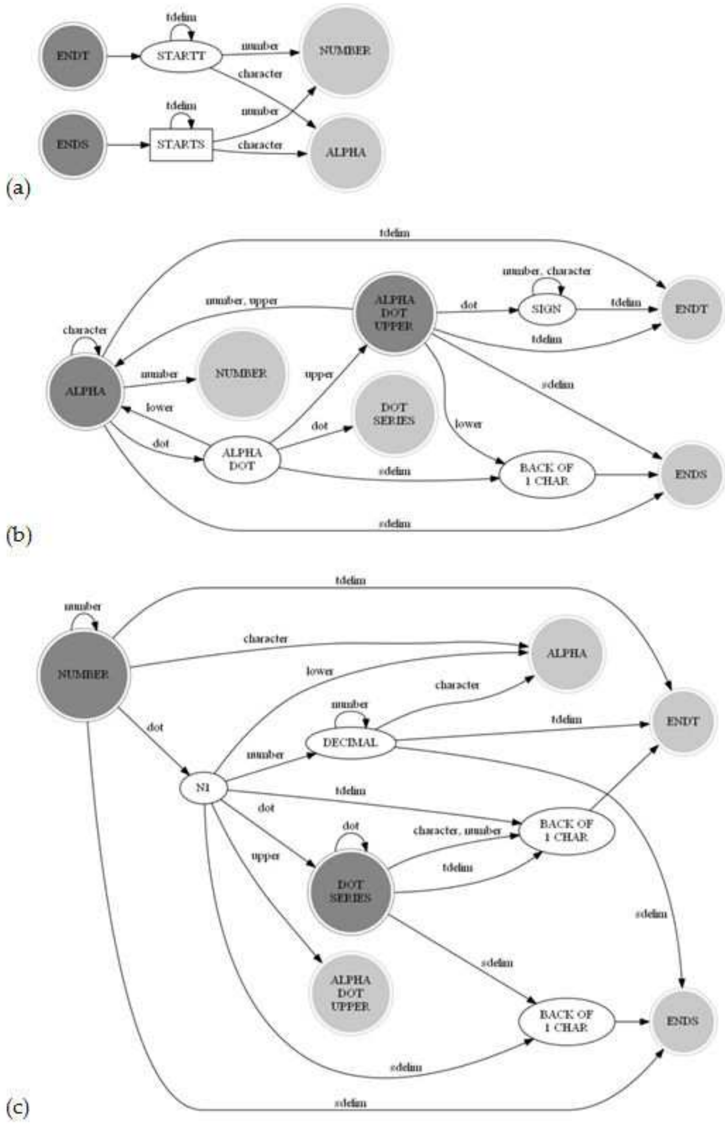


Figure 4. The Finite State Machine that extracts sentences and tokens from text. The three scheme refers to as many sub-procedures

4.2. The abstraction process: from words to domains

The framework uses a semantic network to map tokens into an abstract representation, which can characterize the informative content of the basic textual resource on a cognitive basis. The underlying hypothesis is that to work out the topics addressed in a text, one cannot just depend on the mentioned terms, since each term can in principle convey different senses. On the other hand, the semantic relations that exist between concepts can help understand whether the terms can connect to a single subject or area of interest.

The present approach implements such an abstraction process by mapping tokens into domains. An intermediate step, from tokens to concepts, supports the whole procedure. Two well-known semantic networks have been used to complete this task: EuroWordNet [58], i.e the multilanguage version of WordNet [27], and its extension WordNet Domains [41]. Both EuroWordNet and WordNet Domains are ontologies designed to decorate words or sets of words with semantic relations. The overall structure of EuroWordNet and WordNet Domains are based on the conceptual structures theory [59] which describes the different types of relations that can tie together different concepts.

4.2.1. From tokens to concepts

The abstraction from tokens to concepts is accomplished by using EuroWordNet. EuroWordNet is an extension of WordNet semantic knowledge base for English, inspired by the current sycholinguistic theory of human lexical memory [27]. Nouns, verbs, adjectives and adverbs are organized in sets of synonyms (*synsets*), each of which represents a lexical concept. Actually, the same word can participate in several synsets, as a single word can have different senses (polysemy). Synonym sets are connected to other synsets via a number of semantic relations, which vary based on the type of word (noun, verb, adjective, and adverb); for example, synsets of noun can be characterized by relations such as hyponymy and meronymy. Words can also be connected to other words through lexical relations (e.g., antinomy). EuroWordNet supports different languages; thus, in principle, the approach proposed in this chapter can be easily extended to documents written in Italian, Spanish, French, and German. Table 1 gives, for each language, the number of terms and the number of concepts provided by EuroWordNet [58].

In the present research, the list of concepts that characterize a text is obtained as follows:

a. For each token $t_i \in \Omega$, extract the list of concepts (i.e., synsets) X_i that EuroWordNet associate to the token: $X_i = \{c_k; k = 1,..,N_{c,i}\}$, where $N_{c,i}$ is the number of different concepts in X_i.

b. Assemble the overall list of concepts: $\Sigma = X_1 \cup X_2 \cup X_3 \cup \ldots\ldots.. \cup X_{Nt}$

To not inflate the list of concepts, in this work the tokens that connect to more than eight concepts are discarded. Such threshold has been set empirically by exploiting preliminary experiments. The list of concepts, Σ, represents an intermediate step to work out the domains; this step will be discussed in the next subsection.

The use of synsets to identify concepts possibly brings about the drawback of word disambiguation. The problem of determining which one, out of a set of senses, are invoked in a textual context for a single term is not trivial, and specific techniques [55, 60, 61] have been developed to that purpose. Word disambiguation techniques usually rely on the analysis of the words that lie close to the token itself [61, 62]. Other approaches exploit queries on a knowledge base. A notable example of this approach exploits WordNet Domains and is discussed in [63]. As a matter of fact, word disambiguation methods suffer from both high computational complexity [60, 64] and the dependency on dedicated knowledge bases [65]. In this work, word disambiguation is implicitly obtained by completing the abstraction from concepts to domains.

Language	Number of terms	Number of concepts
English	120160	112641
Italian	37194	44866
Spanish	32166	30350
French	18798	22745
German	17099	15132

Table 1. EuroWordNet: supported languages and corresponding elements

4.2.2. From concepts to domains

WordNet Domains [41] supports the abstraction from concepts to domains. A domain is a structure that gathers different synsets belonging to a common area of interest; thus a domain can connect to synsets that pertain to different syntactic categories. Conversely, one synset can be linked to multiple domains. Each domain groups meanings into homogenous clusters; therefore, one can use the abstraction from concepts to domains to work out the topics that are actually addressed in the underlying set of tokens Ω. This can be done as follows:

a. identify the domains that can be associated to the concepts included in Σ;

b. For each concept $c_{\,l} \in \Sigma$, extract the list of domains Θ_l that WordNet Domains associate to that concept: $\Theta_l = \{d_j; j = 1, \ldots, N_{d,l}\}$, where $N_{d,l}$ is the number of different domains in Θ_l.

c. Obtain the overall list of domains Θ as $\Theta_1 \cup \Theta_2 \cup \Theta_3 \cup \ldots\ldots.. \cup \Theta_{Nc}$, where $N_{\,c}$ is the cardinality of Σ.

design a criterion to work out the foremost domains from Θ.

Different approaches can support the latter step. The implicit goal is to attain word disambiguation, i.e. to remove the ambiguity that may characterize single tokens when they are viewed individually. Thus, one should take advantage of the information obtained from a global analysis; the underlying hypothesis is that the actual topics can be worked out only correlating the information provided by the single tokens. In the present work, that information is conveyed by the list of domains, Θ. The domain-selection algorithm picks out the domains that occur most frequently within the text. The procedure can be formalized as follows:

a. Create an array F with N_d elements, where is the cardinality $|\Theta|$ of set $\Theta = \{d_j; j = 1,..,N_d\}$

b. Set each element of F to 0: $f_j = 0$, $j = 1,.., N_d$

c. For each $t_i \in \Omega$

a. Identify the list of domains to which t_i is linked: $J = \{j \mid d_j \text{ linked to } t_i\}$

b. If $|J| = 1$

$f_j = f_j + 1;$ ▪ $j \in J$

else if $|J| > 1$

$f_j = f_j + 0.5;$ ▪ $j \in J$

The array F eventually measures the relevance of each domain d_j. The algorithm evaluates the relevance of a domain by taking into account the intrinsic semantic properties of a token. Thus, the relative increment in the relevance of a domain is higher when a token can only be linked to one domain. The rationale behind this approach is that these special cases are not affected by ambiguities.

The array of relevancies, F, provides the input to the task designed to work out the most relevant topics and eventually generate the summary.

4.3. Text Summarization

The framework is designed to generate a summary by identifying, in the original text, the textual portions that most correlate with the topics addressed by the document. Two tasks should be completed to attain that goal: first, identifying the topics and, secondly, correlating sentences with the set of topics themselves.

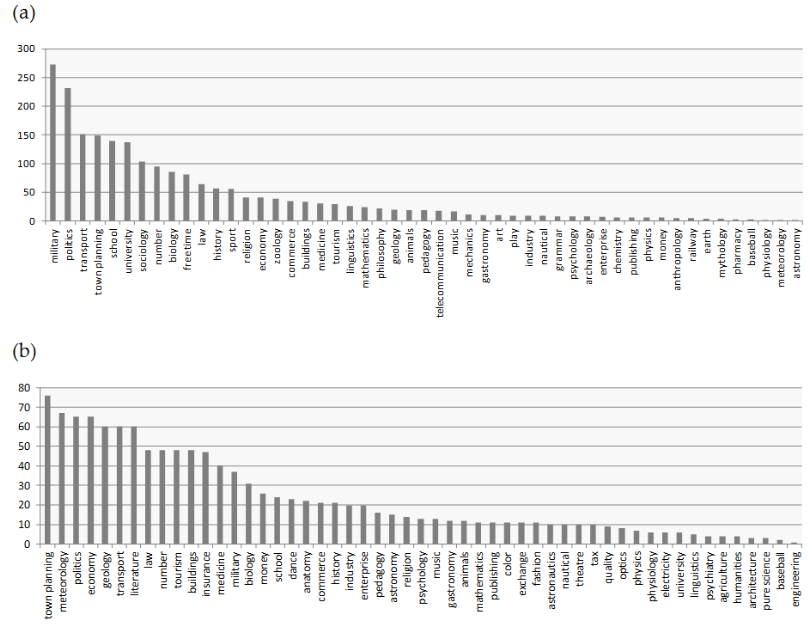


Figure 5. Two examples of array of domains relevancies

The first subtask is accomplished by scanning the array of relevancies, F. In principle, the relevant topics should correspond to the domains having the highest scores in F. However, the distribution of relevancies in the array can play a crucial role, too. Figure 5 illustrates this aspect with two examples. Figure 5(a) refers to a case in which a fairly large gap separates a subset of (highly relevant) domains from the remaining domains. Conversely, Figure 5(b) depicts a case in which the most relevant domains cannot be sharply separated from the remaining domains. The latter case is more challenging as it may correspond either to a text that deals with heterogeneous contents (e.g., the home page of an online newspaper) or to an ineffective characterization of the domains.

To overcome this potential issue, the proposed algorithm operates under the hypothesis that only a limited number of domains compose the subset of relevant topics. The rationale behind this approach is that a tool for content mining is expected to provide a concise description of the Web page, whereas a lengthy list of topics would not help meet such a conciseness constraint. The objective of the algorithm therefore becomes to verify if the array *F* can highlight a limited subset of domains that are actually outstanding.

The algorithm operates as follows. First, a threshold α is used to set a reference value for the relevance score of a domain; as a result, all the domains in *F* that did not achieve the reference value are discarded, i.e., they are considered not relevant. Then, a heuristic pruning procedure is used to further shrink the subset of candidate domains; the eventual goal –as anticipated above- is to work out a limited number of topics.

The selection procedure can be formalized as follows:

a. Sort *F* in descending order, so that f_1 gives the score r_1 of the most relevant domain

b. Obtain F^* by removing from *F* all the domains with relevance smaller than $\alpha\ r_1$

a. If the cardinality of F^* is smaller or equal to θ

b. Else

1. Find the largest gap g_{mn} between consecutive domains in F^*

2. If g_{mn} is larger than χ and *m* is smaller or equal to θ

select as relevant all the domains from d_1 to d_m

3. Else

it is not possible to select relevant domains

The heuristic pruning procedure is applied only if the number of selected domains (i.e., the domains included in F^*) is larger than a threshold θ, which set an upper limit to the list of relevant topics. The heuristic procedure is designed to identify a cluster of relevant domains within the set F^*; to achieve this goal, the gap between consecutive domains is evaluated (the domains in F^* are provided in descending order according to the relevance score). The parameter χ sets the threshold over which a gap is considered significant. As anticipated, the latter procedure may also provide a void subset of relevant topics.

The eventual summary is obtained by picking out the sentences of the original text that most correlate with the relevant topics. To do so, the list of available sentences is sorted in order of relevance scores. Score values are worked out by considering the tokens that form each sentence: if a token can be related to any selected topic, then the relevance of the associate sentence increases. The eventual score of a sentence, finally, stems from normalizing the number of tokens linked to the relevant topics with respect to the total number of tokens that compose the sentence. The procedure can be outlined as follows:

a. Inputs:

The list of selected domains $\Phi = \{d_j; j = 1,..,N_w\}$, where N_w is the cardinality of Φ.

The list of sentences $\Sigma = \{s_l; l = 1,..,N_s\}$, where N_s is the cardinality of Σ.

The list of tokens included in a sentence s_l, $\Omega_l = \{t_{lq}; q = 1,..,N_{tl}\}$, where N_{tl} is the cardinality of Ω_l.

b. Create an array R with N_s elements; each element registers the relevance of the l-th sentence

c. For each sentence $s_l \in \Sigma$

For each token $t_{lq} \in \Omega_l$

If the token can be linked to a domain in Φ

$r_l = r_l + 1$

d. Normalize the elements of R: $r_l = r_l / |\Omega_l|$

The most relevant sentences are obtained by ranking the array R. Actually the selection removes the sentences that are too short to be consistently evaluated. The eventual rank of the sentences is used to build the summary. In general, the summary will include all the sentences that achieved a relevance greater than a threshold.

5. Experimental Results

The DUC 2002 dataset [12] provided the experimental basis for the proposed framework. The dataset has been designed to test methodologies that address fully automatic multi-document summarization. It is organized as follows:

- 59 subjects;
- for each subject, from 5 to 10 different news about that event;
- for each subject, an extractive summary (400 word) created by involving human participants.

Thus, a summarization technique can be evaluated by comparing the outcome of the computer-driven process with that provided by the dataset (the ground truth).

In this work, the DUC 2002 dataset supported two experimental sessions. The first session aimed at evaluating the ability of the proposed framework to generate an effective summary from the documents included in the dataset. The second session was designed to analyze the behavior of the framework in a typical scenario of Web mining: a text source obtained from a Web page that includes different contributions possibly addressing heterogeneous topics.

5.1. The first experimental session: summarization effectiveness

To evaluate the method's ability at effective summarization, this session adopted the ROUGE software [13]. This made it possible to measure the performances of the proposed approach (as per Section 4) on the DUC 2002 dataset.

ROUGE is a software package for automatic evaluation of summaries that has been widely used in recent years to assess the performance of summarization algorithms. The ROUGE tool actually supports different parameterizations; in the present work, ROUGE-1 has been implemented, thus involving 1-gram co-occurrences between the reference and the candidate summarization results. Using DUC 2002 as a benchmark and ROUGE as the evaluation tool allowed a fair comparison between the present approach and other works already published in the literature.

Table 2 gives the results obtained by the proposed framework on the DUC 2002 dataset. The Table compares experiments tested under different configurations of the summarization algorithm; in particular, experimental set-ups differ in the number of sentences used to generate the summary. The first column gives the number of most informative sentences extracted from the original text; the second, third, and fourth columns report on recall, precision, and f-measure, respectively, as measured by ROUGE.

Number of sentences	Recall	Precision	F-measure
10	0.3297	0.5523	0.4028
15	0.4421	0.5747	0.4884
20	0.5317	0.5563	0.5319
25	0.5917	0.5126	0.5382
30	0.6406	0.4765	0.5363

Table 2. The performance achieved by the proposed framework on the DUC 2002 dataset as assessed by ROUGE

Table 2 shows that the methodology presented in this chapter attained results that compared favorably with those achieved by state-of-the-art algorithms [66] on DUC 2002. In this regard, one should consider that the best performance obtained on DUC 2002 is characterized by the following values [66]: recall = 0.47813, precision = 0.45779, F-measure = 0.46729. This confirmed the effectiveness of the underlying cognitive approach, mapping raw text into an abstract representation, where semantic domains identified the main topics addressed within each document. Numerical results point out that the highest F-measure was attained when the summarization algorithm picked out at least the most 20 relevant sentences in a text.

An additional set of experiments further analyzed the outcomes of the proposed approach. In this case, the goal was to understand whether the topic-selection criterion actually fit the criterion implicitly applied by human subjects when summarizing the texts. This involved the array, *F*, measuring the relevance of a set of domains (as per section 4.2.2); for each subject included in DUC 2002, the array *F* was computed with respect to:

- the news linked to that subject;
- the corresponding summary provided by the dataset.

Figure 6 gives a sample of the pair of arrays associated with one of the subjects in the DUC 2002 dataset; in the graph, light-grey lines are associated with the actual reference scores in the benchmark, whereas dark-grey lines refer to the relevance values worked out by the proposed method.

Statistical tools measured the consistency of the domain-selection process: chi-square test runs compared, for each subject, the pair of distributions obtained; the goal was to verify the null hypothesis, namely, that the two distributions came from the same population. The standard value of 0.05 was selected for the confidence level.

The results obtained with the chi-square tests showed that the null hypothesis could *not* be rejected in any of the 49 experiments involved (each subject in DUC 2002 corresponded to one experiment). This confirmed that the distributions of the relevant domains obtained from the whole text could not be distinguished from those obtained from the (human generated) summaries in the DUC 2002 dataset.

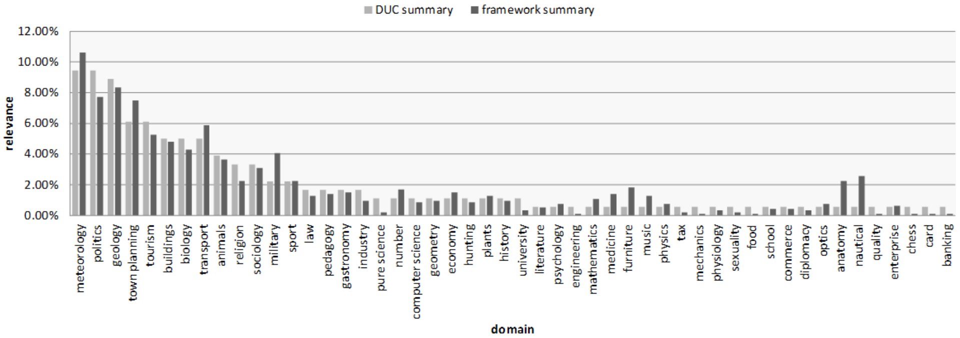


Figure 6. Comparison between the relevance of domains –for the same subject of DUC 2002- in the DUC summary and in the summary provided by the proposed algorithm

5.2. The second experimental session: web mining

The first experimental session proved that the framework can effectively tackle this task (and eventually generate a proper summary) when the input was a news-text, which mainly dealt with a single event. A web page, however, often collects different textual resources, each addressing a specific, homogenous set of topics. Hence, the second experimental session was designed to evaluate the ability of the proposed framework to identify the most informative subsections of a web page.

The experiments involved the DUC 2002 dataset and were organized as follows. A set of new documents were generated by assembling the news originally provided by DUC 2002. Each new document eventually included four news articles and covered four different topics. Then, the list of documents was processed by the proposed framework, which was expected – for each document – to select as the most relevant topics those that were chosen in the set up. Table 3 reports on the results of this experiment; each row represents a single document: the first column gives the topics actually addressed by the document, while the second column gives the topics proposed by the framework. The table reports in boldface the topics that the framework was not able to pinpoint.

Experimental evidence confirmed that the proposed framework yielded satisfactory results in this experiment, too. In this regard, one should also take into account that

- the relative length of the single news somewhat influenced the overall distribution of the topics relevance;
- in several cases the real topics not identified by the framework as the most relevant (i.e., the topics in bold) had relevance scores very close to those characterizing the selected ones.

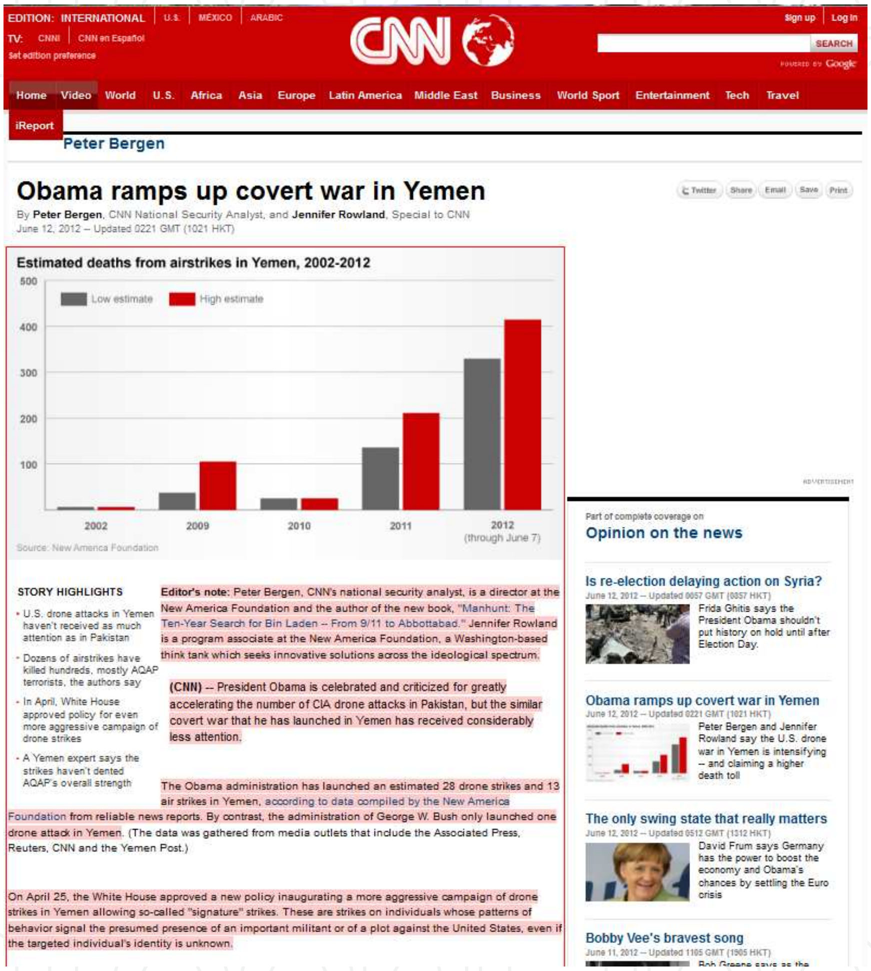
Actual Topics	Topics Proposed by the Framework
Literature / Military / Music / Politics	History / Military / Music / Politics
Literature / **Military** / Music / Politics	Buildings / Literature / Music / Politics
Literature / Military / **Music** / Politics	Literature / Military / Politics / Sociology
Literature / Military / Music / Politics	Biology / Military / Music / Politics
Literature / Military / **Music** / Politics	Military / Politics / School / Sociology
Astronomy / Economy / Music / **Sport**	Astronomy / Biology / Economy / Music
Astronomy / Music / Politics / **Sport**	Biology / Music / Politics / Town Planning
Economy / **Music** / Physics / **Sport**	Economy / Law / Physics / Transport
Music / Physics / Politics / **Sport**	Law / Physics / Politics / Transport
Music / Physics / Politics / Sport	Physics / Politics / Sport / Transport

Table 3. Comparison between actual document topics and topics proposed by the framework

The dataset involved in the experiment was artificially generated to evaluate the effectiveness of the proposed framework in a scenario that resembles a "real word" case. Hence, a fair comparison with other methodologies cannot be proposed. However, Table 3 provides a solid experimental evidence of the efficiency of the approach introduced in this research, as the 'artificial' web pages were composed by using the original news included in the DUC 2002 dataset. As a result, one can conclude that the performances attained by the framework in terms of ability to identify the relevant topics in an heterogeneous document are very promising.

5.3.Web Page Segmentation

The framework can analyze a web page according to two different strategies. The first strategy, identifying the most relevant topics, typically triggers further actions in advanced web-content mining systems: gathering a short summary of the web page (possibly a short summary for each main topic), page segmentation, graphic editing of the web page to favor readability.



Obama ramps up covert war in Yemen

By **Peter Bergen**, CNN National Security Analyst, and **Jennifer Rowland**, Special to CNN
June 12, 2012 -- Updated 0221 GMT (1021 HKT)

Estimated deaths from airstrikes in Yemen, 2002-2012

Source: New America Foundation

STORY HIGHLIGHTS

- U.S. drone attacks in Yemen haven't received as much attention as in Pakistan
- Dozens of airstrikes have killed hundreds, mostly AQAP terrorists, the authors say
- In April, White House approved policy for even more aggressive campaign of drone strikes
- A Yemen expert says the strikes haven't dented AQAP's overall strength

Editor's note: Peter Bergen, CNN's national security analyst, is a director at the New America Foundation and the author of the new book, "Manhunt: The Ten-Year Search for Bin Laden -- From 9/11 to Abbottabad." Jennifer Rowland is a program associate at the New America Foundation, a Washington-based think tank which seeks innovative solutions across the ideological spectrum.

(CNN) -- President Obama is celebrated and criticized for greatly accelerating the number of CIA drone attacks in Pakistan, but the similar covert war that he has launched in Yemen has received considerably less attention.

The Obama administration has launched an estimated 28 drone strikes and 13 air strikes in Yemen, according to data compiled by the New America Foundation from reliable news reports. By contrast, the administration of George W. Bush only launched one drone attack in Yemen. (The data was gathered from media outlets that include the Associated Press, Reuters, CNN and the Yemen Post.)

On April 25, the White House approved a new policy inaugurating a more aggressive campaign of drone strikes in Yemen allowing so-called "signature" strikes. These are strikes on individuals whose patterns of behavior signal the presumed presence of an important militant or of a plot against the United States, even if the targeted individual's identity is unknown.

Part of complete coverage on
Opinion on the news

Is re-election delaying action on Syria?
June 12, 2012 -- Updated 0057 GMT (0857 HKT)
Frida Ghitis says the President Obama shouldn't put history on hold until after Election Day.

Obama ramps up covert war in Yemen
June 12, 2012 -- Updated 0221 GMT (1021 HKT)
Peter Bergen and Jennifer Rowland say the U.S. drone war in Yemen is intensifying -- and claiming a higher death toll

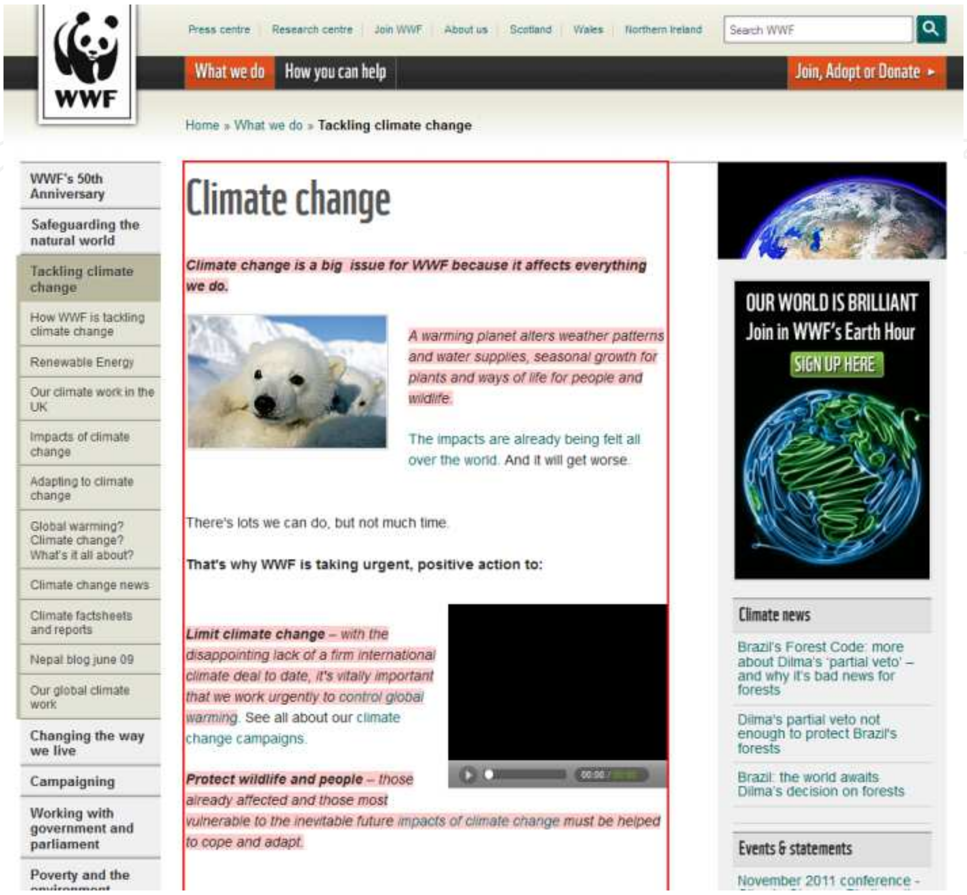
The only swing state that really matters
June 12, 2012 -- Updated 0512 GMT (1312 HKT)
David Frum says Germany has the power to boost the economy and Obama's chances by settling the Euro crisis

Bobby Vee's bravest song
June 11, 2012 -- Updated 1105 GMT (1905 HKT)
Bob Greene says as the

Figure 7. An example of web page analysis supported by the proposed framework

Figure 7 and Figure 8 provide examples of this kind of application. In both cases, the web page included a main section that actually defined the addressed contents, together with other textual parts that did not convey relevant information. The framework supported web content mining by identifying the sentences that actually linked to the relevant topics. These sentences have been highlighted in Figure 7 and Figure 8.

The second strategy typically aims to support users that want to track selected topics. In this case, the goal is to identify the web-page sections that actually deals with the topics of interest. Figure 9 provides an example: the selected topic was ‘pharmacy/medicine,’ and the web page was the ‘News’ section of the publisher *InTech*. The figure shows that an advanced web content mining system could exploit the information provided by the framework to highlight the text parts that were considered correlated with the topic of interest.



Climate change

Climate change is a big issue for WWF because it affects everything we do.

A warming planet alters weather patterns and water supplies, seasonal growth for plants and ways of life for people and wildlife.

The impacts are already being felt all over the world. And it will get worse.

There's lots we can do, but not much time.

That's why WWF is taking urgent, positive action to:

Limit climate change *– with the disappointing lack of a firm international climate deal to date, it's vitally important that we work urgently to control global warming.* See all about our climate change campaigns.

Protect wildlife and people *– those already affected and those most vulnerable to the inevitable future impacts of climate change must be helped to cope and adapt.*

Brazil's Forest Code: more about Dilma's 'partial veto' – and why it's bad news for forests

Dilma's partial veto not enough to protect Brazil's forests

Brazil: the world awaits Dilma's decision on forests

November 2011 conference -

Figure 8. A second example of web page analysis supported by the proposed framework

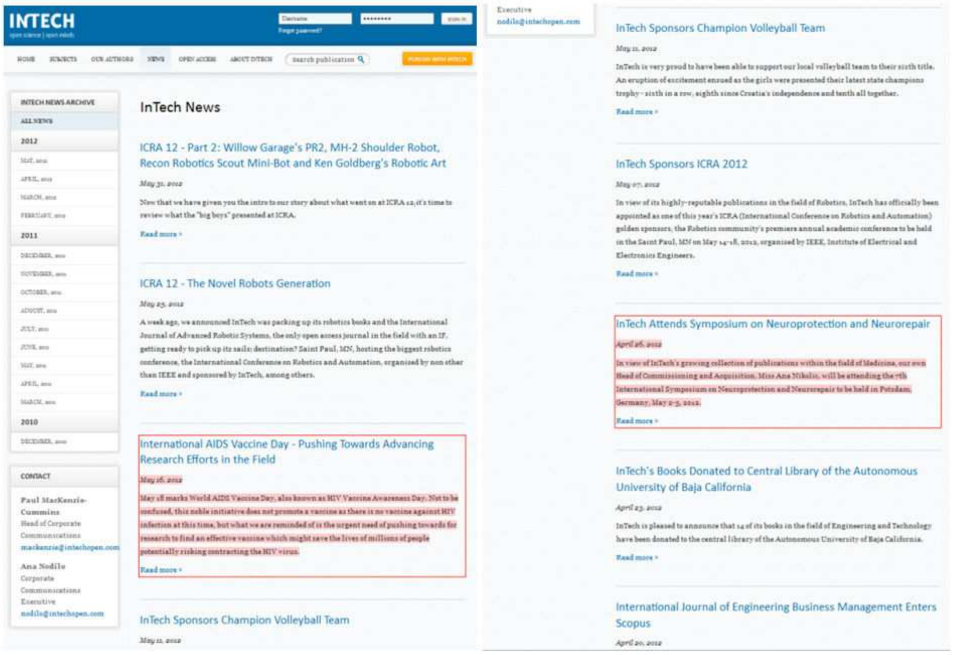
6. Conclusions

The research presented in this chapter introduces a framework that can effectively support advanced Web mining tools. The proposed system addresses the analysis of the textual data provided by a web page and exploits semantic networks to achieve multiple goals: 1) the identification of the most relevant topics; 2) the selection of the sentences that better correlates with a given topic; 3) the automatic summarization of a textual resource. The eventual framework exploits those functionalities to tackle two tasks at the same time: text summarization and page segmentation.

The semantic characterization of text is indeed a core aspect of the proposed methodology, which takes advantage of an abstract representation that expresses the informative content of the basic textual resource on a cognitive basis. The present approach, though, cannot be categorized under the Semantic Web area, as it does not rely on semantic information already embedded into the Web resources.

In the proposed methodology, semantic networks are used to characterize the content of a textual resource according to semantic domains, as opposed to a conventional bag of words. Experimental evidences proved that such an approach can yield a coarse-grained level of sense distinctions, which in turn favors the identification of the topics actually addressed in the Web page. In this regard, experimental results also showed that the system can emulate human assessors in evaluating the relevance of the single sentences that compose a text.

An interesting feature of the present work is that the page segmentation technique is based only on the analysis of the textual part of the Web resource. A future direction of this research can be the integration of the content-driven segmentation approach with conventional segmentation engines, which are more oriented toward the analysis of the inherent structure of the Web page. The resulting framework should be able to combine the outcomes of the two modules to improve the performance of the segmentation procedure.



Paul MacKenzie-Cummins
Head of Corporate Communications
mackenzie@intechopen.com

Ana Nodilo
Corporate Communications Executive
nodilo@intechopen.com

InTech News

ICRA 12 - Part 2: Willow Garage's PR2, MH-2 Shoulder Robot, Recon Robotics Scout Mini-Bot and Ken Goldberg's Robotic Art

May 31, 2012

Now that we have given you the intro to our story about what went on at ICRA 12, it's time to review what the "big boys" presented at ICRA.

Read more »

ICRA 12 - The Novel Robots Generation

May 25, 2012

A week ago, we announced InTech was packing up its robotics books and the International Journal of Advanced Robotic Systems, the only open access journal in the field with an IF, getting ready to pick up its sails; destination? Saint Paul, MN, hosting the biggest robotics conference, the International Conference on Robotics and Automation, organised by non other than IEEE and sponsored by InTech, among others.

Read more »

International AIDS Vaccine Day - Pushing Towards Advancing Research Efforts in the Field

May 16, 2012

May 18 marks World AIDS Vaccine Day, also known as HIV Vaccine Awareness Day. Not to be confused, this noble initiative does not promote a vaccine as there is no vaccine against HIV infection at this time, but what we are reminded of is the urgent need of pushing towards for research to find an effective vaccine which might save the lives of millions of people potentially risking contracting the HIV virus.

Read more »

InTech Sponsors Champion Volleyball Team

May 11, 2012

InTech is very proud to have been able to support our local volleyball team to their sixth title. An eruption of excitement ensued as the girls were presented their latest state champions trophy - sixth in a row, eighth since Croatia's independence and tenth all together.

Read more »

InTech Sponsors ICRA 2012

May 07, 2012

In view of its highly-reputable publications in the field of Robotics, InTech has officially been appointed as one of this year's ICRA (International Conference on Robotics and Automation) golden sponsors, the Robotics community's premiere annual academic conference to be held in the Saint Paul, MN on May 14-18, 2012, organised by IEEE, Institute of Electrical and Electronics Engineers.

Read more »

InTech Attends Symposium on Neuroprotection and Neurorepair

April 26, 2012

In view of InTech's growing collection of publications within the field of Medicine, our own Head of Commissioning and Acquisition, Miss Ana Nikolic, will be attending the 7th International Symposium on Neuroprotection and Neurorepair to be held in Potsdam, Germany, May 2-5, 2012.

Read more »

InTech's Books Donated to Central Library of the Autonomous University of Baja California

April 23, 2012

InTech is pleased to announce that 14 of its books in the field of Engineering and Technology have been donated to the central library of the Autonomous University of Baja California.

Read more »

International Journal of Engineering Business Management Enters Scopus

April 20, 2012

Figure 9. Tracking a selected topic by using the proposed framework

Future works may indeed be focused on the integration of semantic orientation approaches into the proposed framework. These techniques are becoming more and more important in the Web 2.0 scenario, where one may need the automatic analysis of fast-changing web elements like customer reviews and web reputation data. In this regard, the present framework may provide content-filtering features that support the selection of the data to be analyzed.

Author details

Alessio Leoncini, Fabio Sangiacomo, Paolo Gastaldo and Rodolfo Zunino

Department of naval, electric, electronic and telecommunications engineering (DITEN), University of Genoa, Genoa, Italy

References

[1] Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, 2(1), 1-15.

[2] Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., Mcarthur, J., Minton, S., Xheneti, I., Toncheva, A., & Manfrediz, A. (2010). The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through. *Information and Data 2007.*, 1-21.

[3] Naghavi, M., & Sharifi, M. (2012). A Proposed Architecture for Continuous Web Monitoring Through Online Crawling of Blogs. *International Journal of UbiComp*, 3(1), 11-20.

[4] Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 30-40.

[5] Stumme, G., Hotho, A., & Berendt, B. (2006). Semantic Web Mining: State of the art and future directions. *Journal of Web Semantics*, 4(2), 124-143.

[6] Dai, Y., Kakkonen, T., & Sutinen, E. (2011). MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods. *International Journal of Computer Information Systems and Industrial Management Applications*, 3-165.

[7] Thuraisingham, B.M. (2003). Web Data Mining: Technologies and their Applications in Business Intelligence and Counter-terrorism. *Boca Raton: CRC Press*.

[8] Allan, J. (2002). Topic Detection and Tracking: Event-based Information Organization. *Norwell: Kluwer Academic Publisher*.

[9] Chen, H. Discovery of improvised explosive device content in the Dark Web. Proceedings of IEEE International Conference on Intelligence and Security Informatics, ISI ' June (2008). Taipei, Taiwan. 2008. , 08, 17-20.

[10] Yu, S., Cai, D., Wen, J. R., & Ma, W. Y. (2003). Improving pseudo-relevance feedback in web information retrieval using web page segmentation. *Proceedings of the 12th International Conference on World Wide Web, WWW'03,New York, USA.*

[11] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*.

[12] Document understanding conference. (2002). http://www-nlpir.nist.gov/projects/duc/, (accessed 14 May 2012).

[13] Lin, C.Y. Rouge: A package for automatic evaluation of summaries. *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain. 2004.*

[14] Etzioni, O. (1996). The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11), 65-68.

[15] Madria, S. K., Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1999). Research issues in web data mining. *Proceedings of First International Conference on Data Warehousing and Knowledge Discovery, DaWaK'99, Florence, Italy*.

[16] Chakrabarti, S. (2000). Data mining for hypertext. *A tutorial survey*, 1.

[17] Singh, B., & Singh, H. K. Web data mining research: a survey. *Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research, IC-CIC'10. 2010.*

[18] Xu, G., Zhang, Y., & Li, L. (2011). Web Content Mining. *Web Mining and Social Networking*, 6-71.

[19] Liu, B. (2005). Web content mining. *Proceedings of 14th International World Wide Web Conference, WWW'05,May 2005, Chiba, Japan.*

[20] Baumgartner, R., Gottlob, G., & Herzog, M. (2009). Scalable web data extraction for online market intelligence. *Proceedings of the VLDB Endowment*, 2(2), 1512-1523.

[21] Manne, S. (2011). A Novel Approach for Text Categorization of Unorganized data based with Information Extraction. *International Journal on Computer Science and Engineering*, 2846-2854.

[22] Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. *Proceedings of the 15th international conference on World Wide Web, WWW'06,Edinburgh, Scotland*.

[23] Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. *Educational Technology & Society*, 12(4), 30-42.

[24] Maedche, A., & Staab, S. (2001). Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2), 72-79.

[25] Vallet, D., Castells, P., Fernandez, M., Mylonas, P., & Avrithis, Y. (2007). Personalized content retrieval. *context using ontological knowledge. IEEE Transactions on Circuits and Systems for Video Technology 2007*, 17(3), 336-346.

[26] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., & Milios, E. (2006). Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems*, 3(3), 55-73.

[27] Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39-41.

[28] Radev, D. R., Hovy, E., & Mc Keown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, 28(4), 399-408.

[29] Zipf, G. (1949). Human Behaviour and the Principle of Least-Effort. *Cambridge: Addison-Wesley*.

[30] Das, D., & Martins, A. F. T. (2007). A Survey on Automatic Text Summarization. *Engineering and Technology*, 4-192.

[31] Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258-268.

[32] Nenkova, A. (2005). Automatic text summarization of newswire: lessons learned from the document understanding conference. *Proceedings of the 20th national conference on Artificial intelligence, AAAI'05, Pittsburgh, USA.*

[33] García-Hernández, R. A., & Ledeneva, Y. (2009). Word Sequence Models for Single Text Summarization. *Proceedings of the Second International Conferences on Advances in Computer-Human Interactions, ACHI'*, 09, 1-7, *Cancun, Mexico. Washington: IEEE Computer Society; 2009.*

[34] Hennig, L. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. *Proceedings of the Recent Advances in Natural Language Processing Conference, RANLP-2009.*

[35] Svore, K., Vanderwende, L., & Burges, C. (2007). Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL.*

[36] Hannah, M. E., Geetha, T. V., & Mukherjee, S. (2011). Automatic extractive text summarization based on fuzzy logic: a sentence oriented approach. *Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing, SEMCCO'11, Visakhapatnam, India. Berlin: Springer-Verlag*.

[37] Suanmali, L., Salim, N., & Binwahlan, M. S. (2009). Fuzzy Logic Based Method for Improving Text Summarization. *International Journal of Computer Science and Information Security*, 2(1), 65-70.

[38] Barzilay, R., & Elhadad, M. (1997). Using Lexical Chains for Text Summarization. *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization.*

[39] Zamanifar, A., Minaei-Bidgoli, B., & Sharifi, M. (2008). A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text. *Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD'08, Phuket, Thailand.*

[40] Erkan, G., & Radev, D.R. (2004). LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1), 457-479.

[41] Magnini, B., & Cavaglià, G. Integrating Subject Field Codes into WordNet. *Gavrilidou M, Crayannis G, Markantonatu S, Piperidis S, Stainhaouer G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, LREC-2000, 31 May-2 June 2000Athens, Greece.*

[42] Yin, X., & Lee, W. S. (2004). Using link analysis to improve layout on mobile devices. *Proceedings of the Thirteenth International World Wide Web Conference, WWW'04,New York, USA.*

[43] Yin, X., & Lee, W. S. (2005). Understanding the function of web elements for mobile content delivery using random walk models. *Special interest tracks and posters of the 14th international conference on World Wide Web, WWW'05,Chiba, Japan*.

[44] Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). Extracting Content Structure for Web Pages based on Visual Representation. *Proceedings of the 5th Asia-Pacific web conference on Web technologies and applications, APWeb'03, Xian, China. Berlin: Springer-Verlag.*

[45] Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2004). Block-based web search. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'04, Sheffield, UK.*

[46] Ahmadi, H., & Kong, J. (2008). Efficient web browsing on small screens. *Proceedings of the working conference on Advanced visual interfaces, AVI'08, Napoli, Italy.*

[47] Burget, R. (2007). Automatic document structure detection for data integration. *Proceedings of the 10th international conference on Business information systems, BIS'07, Poznan, Poland.*

[48] Burget, R., & Rudolfova, I. (2009). Web page element classification based visual features. *Proceedings of First Asian conference on Intelligent Information and Database Systems, ACIIDS'09, Dong hoi, Quang binh, Vietnam.*

[49] Milic-Frayling, N., & Sommerer, R. (2002). Smartview: Flexible viewing of web page contents. *Poster Proceedings of the Eleventh International World Wide Web Conference, WWW'02,Honolulu, USA.*

[50] Kohlschütter, C., & Nejdl, W. (2008). A densitometric approach to web page segmentation. *Proceeding of the 17th ACM conference on Information and knowledge management, CIKM'08, Napa Valley, USA.*

[51] Cao, J., Mao, B., & Luo, J. (2010). A segmentation method for web page analysis using shrinking and dividing. *International Journal of Parallel, Emergent and Distributed Systems*, 25(2), 93-104.

[52] Borodin, Y., Mahmud, J., Ramakrishnan, I. V., & Stent, A. (2007). The hearsay nonvisual web browser. *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility, W4A'07, Banff, Canada*.

[53] Mahmud, J. U., Borodin, Y., & Ramakrishnan, I. V. (2007). Csurf: a context-driven non-visual web-browser. *Proceedings of the 16th international conference on World Wide Web, WWW'07,Banff, Canada.*

[54] Mehta, R. R., Mitra, P., & Karnick, H. (2005). Extracting semantic structure of web documents using content and visual information. *Special interest tracks and posters of the 14th international conference on World Wide Web, WWW'05,Chiba, Japan*.

[55] Gliozzo, A., Strapparava, C., & Dagan, I. (2009). Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language*, 18(3), 275-299.

[56] libxml: The XML C parser and toolkit of Gnome. http://xmlsoft.org , (accessed 16 May 2012).

[57] Decherchi, S., Gastaldo, P., & Zunino, S. (2009). K-means clustering for content-based document management in intelligence. *Solanas A, Martinez A. (ed.) Advances in Artificial Intelligence for Privacy, Protection, and Security. Singapore: World Scientific*, 287-324.

[58] Vossen, P. (1998). Eurowordnet: A Multilingual Database with Lexical Semantic Networks. *Kluwer Academic Publishers.*

[59] Sowa, J.F. (1992). Conceptual Graphs Summary. *Nagle TE, Nagle JA, Gerholz LL, Eklund PW. (ed.) Conceptual structures. Upper Saddle River: Ellis Horwood*, 3-51.

[60] Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2), 10:1-10:69.

[61] Karov, Y., & Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1), 41-60.

[62] Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), 99-123.

[63] Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The role of domain information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4), 359-373.

[64] Mallery, J.C. (1988). Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. *PhD thesis. MIT Political Science Department Cambridge*.

[65] Gale, W. A., Church, K., & Andyarowsky, D. (1992). A method for disambiguating word senses in a corpus. *Computers and the Humanities*, 26-415.

[66] Barrera, A., & Verma, R. (2011). Automated extractive single-document summarization: beating the baselines with a new approach. *Proceedings of the 2011 ACM Symposium on Applied Computing, SAC'11, TaiChung, Taiwan*.