Using Snippets in Text Summarization: a Comparative Study and an Application

Giuliano Armano, Alessandro Giuliani, and Eloisa Vargiu

Abstract Automatic text summarization consists of automatically creating a summary of one or more texts. As for Web pages, unfortunately classical techniques cannot be applied in presence of dynamic contents. In this paper, we propose the adoption of snippets –i.e., page excerpts provided together with user query results by search engines– as a text summarization technique. The study is conducted along two directions: comparing the proposed approach with a classical text summarization technique and (ii) assessing whether snippet summarization can be successfully applied to contextual advertising. On the one hand, comparative experiments show that the proposed approach has performances similar to those obtained by using the selected classical technique. On the other hand, the adoption of snippets as text summarization technique in contextual advertising show that the performances are quite satisfactory.

1 Introduction

During the 60's, a large amount of scientific papers and books have been digitally stored and made searchable. Due to the limitation of storage capacity, documents were stored, indexed, and made searchable only through their summaries [29]. For this reason, how to automatically create summaries became a primary task and several techniques were defined and developed [18, 12, 25].

More recently, there has been a renewed interest on automatic summarization techniques. The problem now is no longer due to limited storage capacity, but to retrieval and filtering needs. Since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. Therefore, text summarization techniques are currently adopted in sev-

G. Armano, A. Giuliani, and E. Vargiu
University of Cagliari, Dept.of Electrical and Electronic Engineering, Piazza d'Armi, I09123 Cagliari (Italy) e-mail: {armano, alessandro.giuliani, vargiu}@diee.unica.it

eral fields of information retrieval and filtering [7], such as, information extraction [21], text mining [31], document classification [27], recommender systems [23], and contextual advertising [1].

Unfortunately, classical techniques are not easily applicable to dynamic Web pages, which often rely on Microsoft Silverligh[1], Adobe Flash[2], Adobe Shockwave[3], or contain applets written in Java. Conventional parsing methods are often not applicable for the created webpage. Therefore, we claim that snippets, which are provided together with user query results by search engines, might be adopted to perform text summarization on Web pages.

In this paper, we are interested in studying the impact of snippets to perform text summarization. In particular, we conduct the study along two directions: (i) comparing performances obtained by using snippets with those obtained by adopting one of the classical text summarization techniques proposed in [3] and (ii) adopting snippets as text summarization technique in a selected application field, i.e., contextual advertising.

The rest of the paper is organized as follows. Section 2 recalls the main work on text summarization and introduces snippets and their use in search engines. Section 3 presents comparative experiments obtained by adopting snippets with respect to a classical text summarization technique. In Section 4, an application of snippet text summarization in the field of contextual advertising is proposed. Section 5 ends the paper with conclusions and future work.

2 Background

2.1 Text Summarization

Automatic text summarization is a technique in which a text is summarized by a computer program. Given a text, its summary (i.e., a non redundant extract from the original text) is returned.

Mani [19] made a distinction among different kinds of summaries: an *extract* consists entirely of material copied from the input; an *abstract* contains material that is not present in the input or, at least, expresses it in a different way; an *indicative abstract* is aimed at providing a basis for selecting documents for closer study of the full text; an *informative abstract* covers the salient information in the source at some level of detail; and a *critical abstract* evaluates the subject matter of the source document, expressing the abstractor views on the quality of the author's work.

According to [15], summarization techniques can be divided in two groups: those that extract information from the source documents (*extraction-based approaches*) and those that abstract from the source documents (*abstraction-based approaches*).

[1] http://www.microsoft.com/silverlight/

[2] http://www.adobe.com/products/flashplayer.html

[3] http://get.adobe.com/it/shockwave/

The former impose the constraint that a summary uses only components extracted from the source document. These approaches put strong emphasis on the form, aiming to produce a grammatical summary, which usually requires advanced language generation techniques. The latter relax the constraints on how the summary is created. These approaches are mainly concerned with what the summary content should be, usually relying solely on extraction of sentences.

Although potentially more powerful, abstraction-based approaches have been far less popular than their extraction-based counterparts, mainly because generating the latter is easier. An extraction-based summary consists of a subset of words from the original document and its bag of words (BoW) representation can be created by selectively removing a number of features from the original term set. Typically, an extraction-based summary whose length is only 10-15% of the original is likely to lead to a significant feature reduction as well. Many studies suggest that also simple summaries are quite effective in carrying over the relevant information about a document. Straightforward but effective extraction-based text summarization techniques have been proposed and compared in [15]. In a subsequent work, Armano et al. [3] proposed some enriched techniques. In particular, they showed that the technique with best performances in terms of precision, recall, and $F_{measure}$ was the so-called $TFLP$, i.e., the technique that considers the title of the document and its first and last paragraphs.

One may argue that extraction-based approaches are too simple. However, as shown in [9], extraction-based summaries of news articles can be more informative than those resulting from more complex approaches. Also, headline-based article descriptors proved to be effective in determining user's interests [14]. Moreover, these approaches have been successfully applied in the contextual advertising field [5] and in a multimodal scenario [2].

Fig. 1 An example of results given by Yahoo! search engine for the query "Information retrieval".

2.2 Snippets in Search Engines

A general definition of *snippet* is "a small piece of something". In programming, it refers to a small region of reusable source code, machine code, or text. Snippets are often used to clarify the meaning of an otherwise cluttered function, or to minimize the use of repeated code that is common to other functions.

Snippets are also used by search engines to provide a textual excerpt of the corresponding Web page according to the keywords used in the query. Snippet can be considered as a topic-driven summarization, since the summary content depends on the preferences of the user and can be assessed via a query, making the final summary focused on a particular topic. In a preliminary work, Boydell used snippets as summary fragments in the field of social Web [8].

While replying to a user's query, search engines provide a ranked list of related Web pages, each described by a title, a set of snippets, and its URL (see Figure 1). The title is directly taken from the *title* tag of the page, whereas the URL is the *http* address of the page.

For a search engine, the choice of a snippet is an important task. If a snippet shown to the user is not very informative, the user may click on search results that do not contain the information s/he is looking for, or s/he may not click on helpful pages. Moreover, poorly chosen snippets can lead to bad searching experiences. Snippets are usually directly taken from the *description meta* tag, if available. If the description meta tag is not provided, the search engine may use the description for the site supplied by the Open Directory Project (aka, DMoz)[4] or a summary extracted from the main content of the page.

Snippet extraction depends on the adopted search engine. Google[5] does not always use the meta description of the page. In fact, if the content provided by the Web developer in the description meta tag is not helpful, or less than reasonable quality, then Google replaces it with its own description of the site. In so doing, Google snippets will be different, depending on the user's search query. Yahoo![6] provides a patent application that describes how to better decide which snippet to show to users. The gist of Yahoo! patent application is based on three main issues[7]: (i) a query-independent relevance for each line of text, i.e., a degree to which the line of text of the document summarizes the document; (ii) a query-dependent relevance of each of the lines of text, i.e., a relevance of the line of text to the query; and (iii) the intent behind a query. To our best knowledge, Bing[8] developers do not give information on how snippets are extracted. In the literature there are several studies focused on the techniques of snippet extraction, usually relying on algorithms of natural language processing, e.g., as proposed by Li [17].

[4] http://dmoz.org

[5] http://www.google.com

[6] http://www.yahoo.com

[7] http://www.seobythesea.com/2009/12/how-a-search-engine-may-choose-search-snippets/

[8] http://www.bing.com

3 Comparative Study and Results

The first goal of this paper is to compare performances obtained by using snippets with those obtained by adopting a classical text summarization technique. Comparative experiments and the corresponding results are presented in this Section.

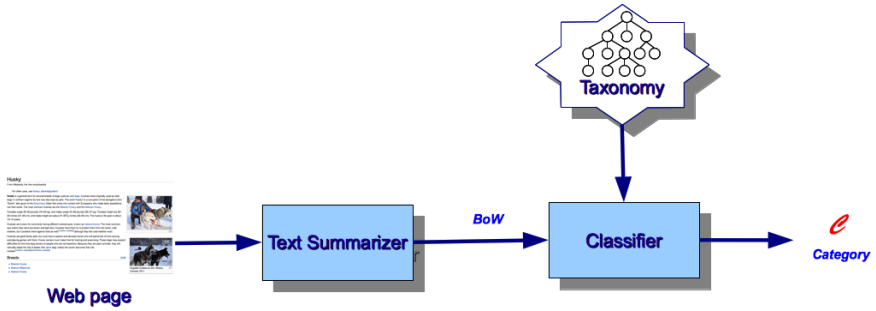

Fig. 2 The system adopted to perform comparative experiments on text summarization.

To perform comparative experiments, we devised a suitable system, depicted in Figure 2, in which the *Text Summarizer* module performs text summarization and the *Classifier* module is a centroid-based classifier aimed at classifying each page in order to calculate precision, recall and $F_{measure}$ of the adopted text summarization techniques. In other words, to assess the text summarization techniques, we used a Rocchio classifier [24] with only positive examples and no relevance feedback, preliminary trained with about 100 Web pages for class. Pages are classified by considering the highest score(s) obtained by the cosine similarity method. To evaluate the effectiveness of the classifier, we performed also a preliminary experiment in which pages are classified without relying on text summarization. The classifier showed a precision of 0.862 and a recall of 0.858.

3.1 Setting Up the Experiments

Experiments have been performed on two datasets extracted by the Open Directory Project and Yahoo! Categories. The former, called BankSearch [28], consists of about 11000 Web pages classified by hand in 11 categories (see Figure 3)[9]. The latter, called Recreation, consists of about 5000 Web pages classified by hand in 18 categories (see Figure 4).

[9] The 11 selected classes are the leaves of the taxonomy, together with the class *Sport*, which contains Web documents from all the sites that were classified as *Sport*, except for the sites that were classified as *Soccer* or *Motor Sport*.

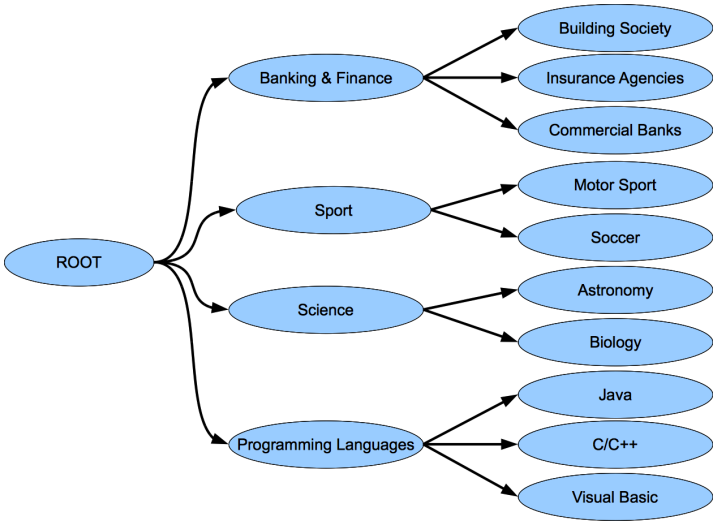

Fig. 3 The taxonomy of BankSearch Dataset.

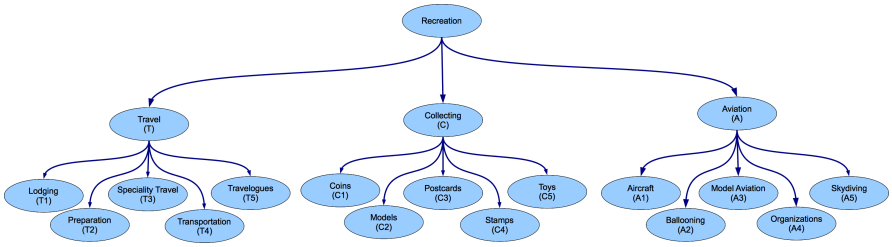

Fig. 4 The taxonomy of Recreation Dataset.

As a baseline for our comparative experiments, we adopted the text summarization technique called *TFLP* (Title, First and Last Paragraph summarization), which considers the title and the first and last paragraphs of the given Web page. This technique, proposed in [3], showed the best results compared with the state-of-the-art techniques proposed in [15]. As for snippets, we performed queries to Yahoo!, asking for the url of each webpage of the dataset, and we used the returned snippets. We performed experiments by considering the snippets by themselves (*S*) and in conjunction with the title of the corresponding Web page (*ST*). It is worth noting that we disregarded dynamic pages from both datasets in order to process the same number of pages independently by the adopted text summarization technique to perform a fair comparison.

3.2 Results

Table 1 reports our experimental results in terms of precision (π), recall (ρ), and $F_{measure}$ (F_1). The Table gives also the average number of extracted terms (T).

The results obtained on BankSearch are better than those obtained on Recreation. Moreover, they point out that, in both datasets, results obtained by relying on snippets together with the title (ST) are comparable with those obtained by adopting $TFLP$. In particular, $TFLP$ performs slightly better in BankSearch, whereas ST performs slightly better in Recreation. This proves that snippets can be adopted as text summarization techniques, especially when classical techniques can not be applied, as in the case of dynamic Web pages.

Let us note that, for each dataset, the average number of terms for the TFLP technique is about twice the number of terms for the method that uses to snippets. This is due to the fact that a snippet is built as a very short text, not less than two rows, wheres in a TFLP summary is usually longer (two complete paragraphs).

Table 1 Results of text summarization techniques comparison.

	BankSearch			**Recreation**		
	TFLP	**S**	**ST**	**TFLP**	**S**	**ST**
π	**0.849**	0,734	0.806	0.575	0.544	**0.595**
ρ	**0.845**	0.730	0.804	**0.556**	0.506	0.554
F_1	**0.847**	0.732	0.805	0.565	0.524	**0.574**
T	26	12	14	26	11	13

4 Using Snippets as Text Summarization Technique in Contextual Advertising

The second goal of this paper is to study the impact of snippet text summarization in a selected application field. Among other relevant information retrieval and filtering fields in which snippet text summarization could be adopted, we concentrate on contextual advertising.

4.1 Contextual Advertising

Web advertising is one of the major sources of income for a large number of websites. Its main goal is to suggest products and services to the ever growing population of Internet users. There are two primary channels for distributing ads: Sponsored Search (or Paid Search Advertising) and Contextual Advertising (or Content Match). Sponsored Search displays ads on the page returned from a search engine

following a query [13]; whereas Contextual Advertising (CA) displays ads within the content of a generic, third party, Web page.

Ribeiro-Neto et al. [22] examined a number of strategies to match pages and ads based on extracted keywords. In a subsequent work, Lacerda et al. [16] proposed a method to learn the impact of individual features using genetic programming. Broder et al. [10] classified both pages and ads into a given taxonomy and matched ads to the page falling into the same node of the taxonomy. Starting from that work, Armano et al. [4] proposed a semantic enrichment by adopting concepts. Furthermore, modern contextual advertising systems use text summarization techniques in conjunction with the model developed in [10] (see, for instance [1, 5]). Since bid phrases are basically search queries, another relevant approach is to view contextual advertising as a problem of query expansion and rewriting [20, 11]. Another perspective consists on addressing a contextual advertising problem as a recommendation task [6]. Thus, authors view the task of suggesting an ad to a Web page as the task of recommending an item (the ad) to a user (the Web page).

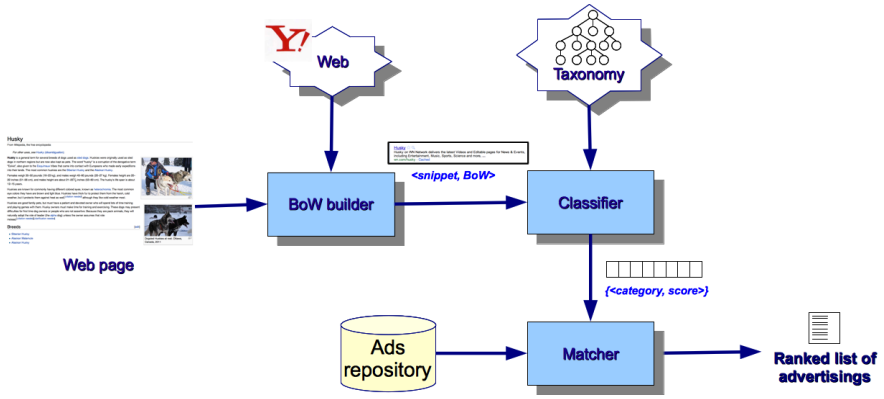

Fig. 5 The implemented contextual advertising system.

4.2 The Implemented System

Being interested in studying the impact of snippets as text summarization technique in contextual advertising, we devised a suitable system (see Figure 5). The system takes a Web page as input. The *BoW builder*, first, retrieves the snippets of the page by asking to Yahoo! search engine and then removes stop-words and performs stemming. This module outputs a vector representation of the original text as *BoW*, each word being represented by its TFIDF [26]. Starting from the *BoW* provided by the *BoW builder*, the *Classifier* classifies the page according to the given taxonomy by adopting a centroid-based approach. This module outputs a vector representation

in terms of Classification Features (CF), each features corresponding to the score given by the classifier to each category. Finally, the *Matcher* ranks the categories according to the scores given by the classifier (i.e., the CF of the target page) and, for each category, randomly extracts a corresponding ad from the *Ads repository*.

Let us note that the proposed system, except for the adopted text summarization technique, is compliant with the system proposed in [1] in which only CF are considered in the matching phase.

4.3 System Performances

To assess the effectiveness of the proposed approach, experiments have been performed on the Recreation dataset described in Section 3.1. As for the ads to be suggested, we built a suitable repository in which ads are classified according to the given taxonomy. In this repository, each ad is represented by the Web page of a product or service company.

Performances have been calculated in terms of *precision at k* with $k \in [1,5]$, i.e., the precision in suggesting k ads. Given a page p and an ad a, the $\langle p,\ a\rangle$ pair has been scored on a 1 to 3 scale defined as follows:

1 - Relevant: a is semantically directly related to the main subject of p, i.e., a and p belongs to the same category;
2 - Somewhat relevant: (i) a is related to a similar subject of p (*sibling*), i.e., a and p belongs to sibling categories; (ii) a is related to the main topic of p in a more general way (*generalization*), i.e., a belongs to the parent node of the category p; or (iii) a is related to the main topic of p in a too specific way (*specification*), i.e., a belongs to a child of the category of p;
3 - Irrelevant. a is unrelated to p, i.e., the category to which a belongs is in a different branch with respect to the category to which p belongs.

According to state-of-the-art contextual advertising systems (e.g., [10]), we considered as True Positives (TP) ads scored as 1 or 2, and a False Positives (FP) ads scored as 3.

Table 2 Precision at k of the proposed contextual advertising system by adopting: $TFLP$ (CA_{TFLP}), the sole snippets (CA_S); and the snippets together with the page title (CA_{ST}).

k	$\mathbf{CA}_{TFLP}$	$\mathbf{CA}_S$	$\mathbf{CA}_{ST}$
1	**0.868**	0.837	0.866
2	0.835	0.801	**0.836**
3	0.770	0.746	**0.775**
4	0.722	0.701	**0.729**
5	0.674	0.657	**0.681**

In performing experiments, we compared the performances obtained by using as text summarization technique: $TFLP$, the resulting system being CA_{TFLP}; the

sole snippets, the resulting system being CA_S; and the snippets together with the page title, the resulting system being CA_{ST}. Let us note that, as the focus of this paper is on text summarization, comparative experiments among the implemented contextual advertising system and selected state-of-the-art systems are out of the scope of this work. Nevertheless, let us stress that CA_{TFLP} coincides with the system proposed in [5] in which the α parameter is set to 0 (i.e., only CF are considered in the matching phase).

Table 2 shows that, for all the compared systems, results are quite satisfactory, especially in suggesting 1 or 2 ads. It also clearly shows that, except for $k = 1$, CA_{ST} is the system that performs better. This proves the effectiveness of adopting snippets as text summarization technique in the field of contextual advertising.

5 Conclusions and Future Work

Since classical text summarization techniques are not applicable for dynamic Web pages, in this paper we proposed to use snippets. The aim of the paper was twofold: (i) to compare performances obtained by using snippets with those obtained by adopting a classical text summarization technique and (ii) to study the impact of snippets in a selected application field, i.e., contextual advertising. The comparisons showed that the proposed snippet text summarization technique has performances (in terms of precision, recall, and F_1) similar to those obtained by using a classical technique (i.e., $TFLP$). The adoption of snippets as text summarization technique in contextual advertising showed that performances, calculated in terms of precision at k, are quite good, especially in suggesting 1 or 2 ads, and that the system that uses both snippets and title is the one with the best performances.

As for future work we are planning to perform further comparative experiments with the methods described in [18, 30, 12].

Acknowledgment

This work has been partially supported by Hoplo srl. We wish to thank, in particular, Ferdinando Licheri and Roberto Murgia for their help and useful suggestions.

References

1. Anagnostopoulos, A., Broder, A.Z., Gabrilovich, E., Josifovski, V., Riedel, L.: Just-in-time contextual advertising. In: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 331–340. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1321440.1321488

2. Armano, G., Giuliani, A., Messina, A., Montagnuolo, M., Vargiu, E.: Experimenting text summarization on multimodal aggregation. In: 5th International Workshop DART 2011, New Challenges on Information Retrieval and Filtering, CEUR Workshop Proceedings, Vol. 771. C. Lai and G. Semeraro and E. Vargiu (2011)
3. Armano, G., Giuliani, A., Vargiu, E.: Experimenting text summarization techniques for contextual advertising. In: IIR'11: Proceedings of the 2nd Italian Information Retrieval (IIR) Workshop (2011)
4. Armano, G., Giuliani, A., Vargiu, E.: Semantic enrichment of contextual advertising by using concepts. In: International Conference on Knowledge Discovery and Information Retrieval (2011)
5. Armano, G., Giuliani, A., Vargiu, E.: Studying the impact of text summarization on contextual advertising. In: 8th International Workshop on Text-based Information Retrieval (2011)
6. Armano, G., Vargiu, E.: A unifying view of contextual advertising and recommender systems. In: Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010), pp. 463–466 (2010)
7. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
8. Boydell, O., Smyth, B.: From social bookmarking to social summarization: an experiment in community-based summary generation. In: Proceedings of the 12th international conference on Intelligent user interfaces, IUI '07, pp. 42–51. ACM, New York, NY, USA (2007)
9. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. Inf. Process. Manage. **31**, 675–685 (1995)
10. Broder, A., Fontoura, M., Josifovski, V., Riedel, L.: A semantic approach to contextual advertising. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 559–566. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1277741.1277837
11. Ciaramita, M., Murdock, V., Plachouras, V.: Online learning from click data for sponsored search. In: Proceeding of the 17th international conference on World Wide Web, WWW '08, pp. 227–236. ACM, New York, NY, USA (2008)
12. Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**, 264–285 (1969)
13. Feldman, J., Muthukrishnan, S.: Algorithmic methods for sponsored search advertising. CoRR **abs/0805.1759** (2008)
14. Kołcz, A., Alspector, J.: Asymmetric missing-data problems: Overcoming the lack of negative data in preference ranking. Inf. Retr. **5**, 5–40 (2002)
15. Kolcz, A., Prabakarmurthi, V., Kalita, J.: Summarization as feature selection for text categorization. In: CIKM '01: Proceedings of the tenth international conference on Information and knowledge management, pp. 365–370. ACM, New York, NY, USA (2001)
16. Lacerda, A., Cristo, M., Gonçalves, M.A., Fan, W., Ziviani, N., Ribeiro-Neto, B.: Learning to advertise. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 549–556. ACM, New York, NY, USA (2006). DOI http://doi.acm.org/10.1145/1148170.1148265
17. Li, Q., Chen, Y.P.: Personalized text snippet extraction using statistical language models. Pattern Recogn. **43**, 378–386 (2010)
18. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development **2**(2), 159–165 (1958)
19. Mani, I.: Automatic summarization. John Benjamins, Amsterdam (2001)
20. Murdock, V., Ciaramita, M., Plachouras, V.: A noisy-channel approach to contextual advertising. In: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising, ADKDD '07, pp. 21–27. ACM, New York, NY, USA (2007)
21. Rau, L.F., Jacobs, P.S., Zernik, U.: Information extraction and text summarization using linguistic knowledge acquisition. Inf. Process. Manage. **25**, 419–428 (1989)
22. Ribeiro-Neto, B., Cristo, M., Golgher, P.B., Silva de Moura, E.: Impedance coupling in content-targeted advertising. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 496–503. ACM, New York, NY, USA (2005). DOI http://doi.acm.org/10.1145/1076034.1076119

23. Ricci, F., Rokach, L., Shapira, B., Kantor, P.: Recommender Systems Handbook. Springer, US (2010)
24. Rocchio, J.: The SMART Retrieval System: Experiments in Automatic Document Processing, chap. Relevance feedback in information retrieval, pp. 313–323. PrenticeHall (1971)
25. Salton, G., Buckley, C.: On the use of spreading activation methods in automatic information. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '88, pp. 147–160. ACM, New York, NY, USA (1988)
26. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1984)
27. Shen, D., Chen, Z., Yang, Q., Zeng, H.J., Zhang, B., Lu, Y., Ma, W.Y.: Web-page classification through summarization. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04, pp. 242–249. ACM, New York, NY, USA (2004)
28. Sinka, M., Corne, D.: A large benchmark dataset for web document clustering. In: Soft Computing Systems: Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications, pp. 881–890. Press (2002)
29. de Smedt, K., Liseth, A., Hassel, M., Dalianis, H.: How short is good? An evaluation of automatic summarization, pp. 267–287. Museum Tusculanums Forlag, Kbenhavn (2005)
30. Tsegay, Y., Puglisi, S.J., Turpin, A., Zobel, J.: Document compaction for efficient query biased snippet generation. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR '09, pp. 509–520. Springer-Verlag, Berlin, Heidelberg (2009)
31. Witten, I.H., Bray, Z., Mahoui, M., Teahan, B.: Text mining: A new frontier for lossless compression. In: Proceedings of the Conference on Data Compression, DCC '99, pp. 198–. IEEE Computer Society, Washington, DC, USA (1999)