An Exploratory Study of Navigating Wikipedia Semantically: Model and Application

I-Chin Wu[1], Yi-Sheng Lin[1], and Che-Hung Liu[2]

[1] Department of Information Management, Fu-Jen Catholic University
No.510, Zhongzheng Rd., Xinzhuang Dist., New Taipei City 24205 Taiwan
[2] Department of Business and Management, National University of Tainan
33, Sec. 2, Su-Lin St. Tainan, 700, Taiwan
icwu.fju@gmail.com, chehung@mail.nutn.edu.tw

Abstract. Due to the popularity of link-based applications like Wikipedia, one of the most important issues in online research is how to alleviate information overload on the World Wide Web (WWW) and facilitate effective information-seeking. To address the problem, we propose a semantically-based navigation application that is based on the theories and techniques of link mining, semantic relatedness analysis and text summarization. Our goal is to develop an application that assists users in efficiently finding the related subtopics for a seed query and then quickly checking the content of articles. We establish a topic network by analyzing the internal links of Wikipedia and applying the *Normalized Google Distance* algorithm in order to quantify the strength of the semantic relationships between articles via key terms. To help users explore and read topic-related articles, we propose a SNA-based summarization approach to summarize articles. To visualize the topic network more efficiently, we develop a semantically-based WikiMap to help users navigate Wikipedia effectively.

Keywords: Navigation, Normalized Google Distance, Semantically-based, SNA-based summary, Wikipedia.

1 Introduction

With the ubiquity of the Internet and Web 2.0 technologies, the WWW has become the main source of information and knowledge in the modern era. The top sites on the Web, as ordered by Alexa traffic rank, are Google, Facebook, YouTube, Yahoo, Live, Baidu, Wikipedia, Blogger, MSN, and Tencent (October, 2010). Wikipedia is the most popular web-based, free-content encyclopedia web site out of the top 10 Web sites. The statistical data from Wikipedia shows that in 2008, the site welcomed 684 million visitors; there were more than 91,000 contributors working on more than 16 million articles. Because of the popularity of applications like Wikipedia, the number of articles in Wikipedia is constantly expanding. As we know, more and more people regard Wikipedia is an efficient means to find needed knowledge, such as searching definitions of terminologies, exploring articles on related topics, and so on. Basically, Wikipedia

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, HCII 2011, LNCS 6778, pp. 140–149, 2011.

users browse content in the traditional manner (i.e., by following hyperlinks) when searching for information. However, users may unconsciously change their search goals or get lost when exploring or retrieving information in Wikipedia. In order to make research more effective for the vast number of Wikipedia users, it is highly important to develop effective search or navigation tools to guide users to find and organize needed information or topics.

Generally, users invest a great deal of time browsing by following links or searching for specific information. Because of the rapid growth in the volume of information on the WWW, web mining and information retrieval are regarded as key techniques for finding desired information. Web mining tries to extract potentially useful implicit information, link structures and patterns from information units or activities on the WWW. There are three types of web mining techniques: web content mining, web structure mining, and web usage mining. The main difference between web pages and static text documents is that the former contain content as well as link information, and metadata [1][9]. Web content mining exploits IR and artificial intelligence (AI) techniques to extract, mine and analyze information from web pages. Generally, web content mining strategies can be divided into those implicitly that mine information or knowledge from the content of documents, and those designed to improve the search results, i.e., the information retrieved by search engines. IR technology relies primarily on content analysis techniques, but Web pages are usually noisy and contain various types of content, such as text, images, and multimedia. To resolve this problem, some researchers have exploited the hyperlink structure, which provides hyperlink information for a collection of web pages, and proposed ranking algorithms to rank search results. Analyzing the hyperlink structure between WWW pages to support user search activities has attracted a great deal of attention in recent years. Since the link structure encodes a considerable number of latent human judgments, link mining and analysis techniques are employed by commercial search engines, e.g., the *PageRank* algorithm [2] used by the Google search engine is one of the most well-known link-based algorithms. Almpanidisa and Kotropoulos (2007) **[1]** proposed a topical information resource discovery algorithm that implements a focused or topic-driven crawler by combining text and link analysis techniques. Their results show that, in the initial stage, the content-based and link-based algorithm does not need a lot of data and that it outperforms comparable methods.

In Wikipedia, a topic may contain many articles; thus, it is difficult for users to locate articles relevant to the topic simply by following the hyperlinks in the articles. To address this problem, we propose a semantically-based navigation system that is based on the theories and techniques of link mining, semantic relatedness analysis, social network analysis (***SNA***) and text summarization. Specifically, we employ a *Link Strength* (***LS***) measure to establish a preliminary topic network by analyzing Wikipedia's internal links [13]. Our goal is to find the specific topic or related subtopics for a seed query (topic) to construct a preliminary internal link-based network. Moreover, we refine the *Normalized Google Distance* algorithm [3] in order to quantify the strength of the semantic relationships between articles via key terms, and filter out articles that do not have strong semantic relationships. Our preliminary evaluation results demonstrate the effectiveness of applying semantic analysis in an internal link-based network. To help users search for information, we apply centrality-based and cohesive measures in ***SNA*** to summarize multiple articles. The measures

are *k-clique* and *degree centrality,* which identify the sub topics of the seed query and key articles of the sub topics respectively. Then, when the user clicks on a topic node that he/she wants to explore, an SNA-based summary is presented on the interface. To more efficiently visualize the semantically-based topic network, an interface as it would appear on PCs and mobile devices is developed to help users navigate Wikipedia effectively.

2 Basic Concepts

2.1 Semantic Relatedness Analysis: Normalized Google Distance

Humans acquire the meaning of words and the relationship among words according to their background knowledge. For example, many humans could answer the question of how "polar bears" or "automobiles" are related to "global warming." However, it is difficult for computers to make judgments regarding the semantic relationships between keywords. How to find the correct tasks to allow computers to automatically extract the meaning of words has recently gained a great deal of attention among the natural language processing and artificial intelligence communities [3] .

Basically, there are mainly three measures to estimate the semantic relatedness of different words, which are thesaurus-based, corpus-based and Wikipeida-based measures [5]. Most of the research methods rely on long-term and labor-intensive efforts to construct the semantic relationships among words. Recently, an automatic algorithm has been proposed using the search results counts from the Google search engine, i.e. the *normalized Google distance* (***NGD***) algorithm [3]. The ***NGD*** algorithm can detect the semantic relationship among terms using the Google search engine. Based on reports from WorldWideWebSize.com, the Google search engine indexes 23.6 billion pages (December, 2010). Thus, the WWW is the largest corpus that can be used to analyze the semantic relationships among words. The main idea of the ***NGD*** is to understand the relationships between any two terms according to the number of search results, i.e., the number of return pages. Vitanyi and Cilibrasi (2007) [3] provided a statistical index based on Google page counts, showing the logical distance of a pair of terms called ***NGD***. When the value gets lower, it implies there is a closer relation between two terms. Eq. (1) is the normalized Google distance based on conditional probability. That is, the probability p(x|y) is defined by p(x|y)=p(x,y)/p(y) and $p(y) = y/M$. In the *NGD* equation, *y* means the probability, as the number of words that Google searched. The result number, by which it is divided, is ***M*** (total numbers of document indexed by Google search). *p*(x, y) is the probability of queries when searching two terms at the same time. Furthermore, because the conditional probabilities are independent of ***M***, we use frequency, i.e., number of search pages, instead of conditional probability to derive the Eq. (2).

$$D(x,y) = \frac{\max\{\log 1/p(x|y), \log 1/p(y|x)\}}{\max\{\log 1/p(x), \log 1/p(y)\}} \tag{1}$$

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \tag{2}$$

2.2 Automatic Summarization Techniques

With the fierce growth of information on the Web, it is important to extract the key abstract content from information sources. Therefore, text summarization techniques and tools are useful to help users find needed information and make decision quickly. Basically, the aim of text summarization is to extract key sentences from one or more documents to represent the meaning of the target document(s). Generally, text summarization can be divided into three phases: analyzing the source text, determining the salient points and synthesizing an appropriate output [7].

For the number of documents, it could be classified as single-document summarization and multi-document summarization. Single-document summarization is generated from the content of one document by different methods that aim to reduce the redundant information. Forsyth and Rada (1996) [6] researched the approach of computing the TF-IDF weight of a document to figure out important terms. Based on their research, they found that sentences are composed from many terms; thus, they attempt to extract important sentences using signature words of the document. Teufel and Moens (1997) [11] employed five heuristic methods: cue, location sentence length, thematic word method, and title methods to extract important sentences from document training sets. They examine the effectiveness of each method and then integrated five methods to analyze the tradeoff between methods. The experiment results show that cue phrases (i.e. in summary, in conclusion, in short, therefore or proper nouns) are the strongest single heuristic and the combination of 5 heuristics will lead to the best performance. Hovy and Lin (1997) [8] mainly considered the important position of sentences (i.e. the first sentence of a document or a phrase), called sentence position, to generate the summarization directly.

3 The System Framework

With the emergence of Web 2.0 technologies, social web sites (i.e., social networking websites and micro-blogging services) provide unprecedented opportunities for sharing user-generated content. Wikipedia, one of the most famous collaborative projects on the Web, has become an extremely popular reference database for people seeking information or knowledge. The process for generating a semantically-based topic network is illustrated in Fig. 1. In the following, we describe the modules of the framework.

Article pre-processing module. The proposed framework retrieves all link-related articles within three degrees. Besides extracting the hyperlinks from each article, the module stores the content of the article, including the title and its associated links, in XML format. Basically, the Wikipedia is written in wiki markup language which is based on the XML web language composed of wiki-recognized title, content, and most of all containing huge numbers of out-links and in-links. In this phase, we extract and indentify information we want to use from those special symbols of wiki language. For example, the symbols “[” and “]” are used for out-links, while the symbols “[[” and “]]” present the in-links.

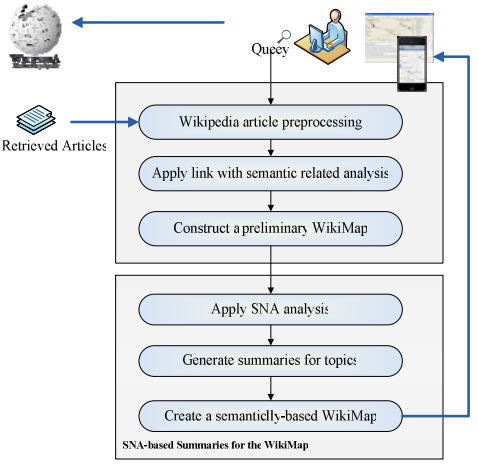

Fig. 1. The process for generating an SNA-based summaries interface

Internal link-based and semantic network analysis module. This module utilizes the proposed link strength (***LS***) measure to search for topics related to the user's seed query. It filters out unimportant articles and finds possible subgroups around the seed query. If the value or strength of a link is less than a specified threshold, the system will remove the link from the internal link-based network. Then, the system executes semantic analysis based on the ***NGD*** algorithm to filter noise nodes from the initial network.

SNA analysis module. This module applies the social network analysis (***SNA***) indicators defined in social network theory [12] to recognize the role of articles in the topic network based on the results of the previous stage. Basically, we mainly apply centrality and cohesive measures in SNA, i.e., k-cliques and degree centrality, to identify subtopics and key articles of subtopics to help users effectively search for information in Wikipedia.

SNA-based summaries generation module. The aim of this step is to generate the summaries of articles based on the SNA analysis results, namely, the interface will present different summarization results based on the topologies of the network. To help users search for information, we apply ***SNA*** indicators to summarize single and multiple articles. Then, when the user clicks on a topic node that he/she wants to explore, an SNA-based summary is presented on the interface to help the user quickly read the related articles.

4 Incorporating Semantic Analysis into the Internal Link-Based Network

4.1 Internal Link Analysis by the LS Measure

We use the term "article" to denote an entry in Wikipedia rather than a page on the WWW, and the term "node" to denote a word in an article with a hyperlink to another

article. The link strength (***LS***), which indicates the degree of closeness between two articles, is determined by considering the type and frequency of the links between the articles. Our goal is to find the specific topic or related subtopics for a *seed query*. An article may have three types of links: in-links, out-links, and reciprocal bi-directional or multi-directional links. Two articles linked to each other by the same nodes have a bidirectional link. In addition to the relationships between articles, the frequency of the links (the link frequency) between the nodes is determined by the *LS measure,* denoted by **ζ,** which is calculated as follows:

$$\zeta(a_i, a_j) = (f\_bi(a_i, a_j))^{w1} + w_2 \times f\_in(a_i, a_j) + w_3 \times f\_out(a_i, a_j), \quad (3)$$

where $f\_in(a_i, a_j)$ denotes the frequency of in-links from a_j to a_i; $f\_out(a_i, a_j)$ denotes the frequency of out-links from a_i to a_j ; and $f\_bi(a_i, a_j)$ denotes the frequency of bidirectional links between a_i and a_j. Details of setting the relative weights of in-links and out-links and the threshold of the *LS* value is shown in our recently work [13].

4.2 Semantic Relatedness Analysis by the NGD Algorithm

For filtering articles with low semantic relatedness in the initial internal link-based network, we conduct further tests for semantic relatedness via the key terms of each article. As we know, it is a difficult task to make judgments of the semantic relationships between keywords via computers. Recently, an automatic algorithm has been proposed using the number of search results returned by the Google search engine, i.e. normalized Google distance (***NGD***) algorithm, as we introduced in Section 2.1. In our research, we defined the titles of the articles in Wikipedia as the nodes in the ***NGD*** algorithm. We can find out the relevancy strength of two web pages by calculating and analyzing the titles of these two pages. A value will be given to the distance of the relationship using the ***NGD*** algorithm. Notably, the lower the value is, the higher the semantic relationship is between articles. We filter topics that have a distance value higher than the threshold aim in order to remove noise. The formula of the algorithm is given in Equation (2).

Table 1. ***NGD*** threshold value of the seed query of "History of personal computers"

Title of the Article	Judgment(Irrelevant)	NGD Value
74181(ALU)	2	0.464
BASIC	2	0.876
Bluetooth	2	0.971
Burroughs_Corporation	3	0.444
Laser_diode	2	1.000
Left-handedness	3	1.000
Pixar	2	0.896
PlayStation_2	3	0.736
The_Walt_Disney_Company	3	0.421
Video_game	3	0.836
Wii_Remote	5	0.421
		Average: 0.695

Testing of *NGD* threshold. Based on Evangelista and Kjos-Hanssen's research (2006) [4], the expected value of the ***NGD*** threshold will be around 0.7. Accordingly, we invited experts from the information management department to determine if the nodes in the topic network are relevant to the seed query of "History of the personal computer". Table 1 shows the results of the unrelated nodes' titles with the seed query, derived using the judgment of the experts. The results shows that the average ***NGD*** value of all those unrelated nodes is 0.695, which is similar to the execrated results from previous research, i.e., the ***NGD*** value is 0.7 [4]. In our the other test, the average ***NGD*** value of all those unrelated nodes for the seed query of "knowledge management" was 0.732. Thus, we set the value of the threshold as 0.7 to construct the final topic network toward the seed query.

Similarity Calculation. Notably, we also conduct a similarity calculation using the cosine measure to investigate the cohesiveness of each network. Table 2 shows that after applying ***NGD*** filtering, the networks will have a higher cosine value than those for which semantic analysis has not been conducted.

Table 2. Similarity comparison between IIN and IIN with NGD

Queries	Internal-link based network (IIN)	IIN With NGD Analysis
History of PCs	**0.248**	0.240
Star Trek	0.149	**0.275**
Abraham Lincoln	0.103	**0.148**
Knowledge Management	0.261	**0.291**

5 SNA-Based Summaries for the Topic Network

5.1 Process for Generating Summarization

Summaries can be divided into three types based on purpose: indicative summaries, informative summaries and critical summaries. In our research context, we generate informative summaries to help users to explore the topics in Wikipedia, with the aid of shorter versions of the original articles. The information summary provides a highly reduced list of the important content of the document, which users could even use to replace the original document. The 30% of a text that is important always represents 80% to 90% of the article's original focus. Fig. 2 shows the two phases needed to generate the **SNA**-based summaries, i.e., the analysis phase and the synthesis phase, in our research context. The steps for generating summarization proceeded as follows: (1) Pre-process the source articles; (2) parse articles into terms; (3) select the terms based on the feature set; (4) weight the terms based the role of the articles in the SNA; (5) Calculate sentence scores based on the feature set; and (6) generate summaries from top-N sentences and then resort them based on the sentences' original order in the articles.

We consider the roles of articles in the social network. Accordingly, we then apply different summarization strategies based on the role of articles in the social network. First, the cohesive article (***CA***) will be identified based on the formula of *k-cliques* given in the social network analysis (***SNA***). The analysis results help us label the sub-

topics of the nodes in the network. Second, we will find the hub article (***HA***) of each sub-topic. The ***HA*** is analyzed based on the formula of *degree centrality* given in ***SNA***. This seeks to help the user explore the main topic of each sub-topic. Similarly, we also present a guided summarization of the hub article and its associated articles.

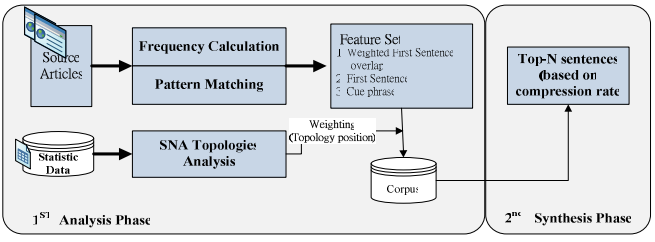

Fig. 2. Process for generating SNA-based summaries

5.2 Identify Sub-topics and Extract the Summaries for the Topics

Radev et al. (2004) [10] presented the MEDA algorithm to generate the summaries of multiple multi-documents. The algorithm adopted three features to select the top sentences which are centroid value, positional value, and first-sentence overlap. Similar to previous researchers, we consider the characteristics of the articles in Wikipedia in order to extract key terms and then select the top sentences for extracting summaries of the topic or sub topic. In this work, we proposed three features, which are first sentence, weighted first-sentence overlap and cue phrases to select key terms. Particularly, we incorporated the concept of the position value of a sentence into the concept of the first-sentence overlap to generate the score of each sentence, i.e. weighted first-sentence overlap, wf_i, as shown in Eq.(4). Basically, *position value,* P_i, means the position of the ith sentence in the document, and n is the number of sentences of the target article. Furthermore, the candidate terms can be selected from these sentences, i.e., sentences with high value of wf_i, as shown in Eq. (5). The final key terms are selected based on different relative importance of the first sentence $\overrightarrow{S_1}$, weighted first-sentence overlap $\overrightarrow{a_k}$, and cue phrases, $\overrightarrow{CP_k}$,as shown in Eq.(6).

$$wf_i = Sim(\overrightarrow{f_1}, \overrightarrow{f_i}) \times P_i$$
$$where\ P_i = \frac{n-i+1}{n} \tag{4}$$

$$\overrightarrow{a_k} = \sum_{i=2}^{n} wf_{s_i} \times \overrightarrow{S_i} \tag{5}$$

$$\overrightarrow{a'_k} = \lambda \times top(\overrightarrow{a_k}) + (1-\lambda) \times (\overrightarrow{S_1} + \overrightarrow{CP_k}) \tag{6}$$

6 The Illustrative Example and Application

The following table shows the results of the article summarization for the seed query "knowledge management." We adjusted the relative importance of the weights for the weighted first-sentence overlap, first sentence, and cue phrases based on the Eq. (6). We asked two experts to select the sentences which are highly relevant and useful to assist novices understand the subject of "knowledge management." There are 20 distinct sentences are selected from the article for we set the value of λ at 0.1, 0.5, and 0.9 respectively, based the Eq. (6). Both of experts agree that the meaningful and useful sentences will be extracted when we set λ at 0.5. Table 3 shows the results of the top-5 sentences selected from the article for we set the value of λ at 0.5. The sentences with bold type denote the overlapping sentences among three different settings of weights. Finally, we present the interface for the semantically-based WikiMap as it would appear on a mobile device and personal computer, as shown in Figure 3(a) and 3(b) respectively. In the future, the effectiveness of the proposed applications will be evaluated in the real setting.

Table 3. Article summarization for the seed query "knowledge management"

The weight of weighted *first-sentence overlap is set* to 0.5
1. Knowledge Management efforts typically focus on organizational objectives such as improved performance, competitive advantage, innovation, the sharing of lessons learned, integration and continuous improvement of the organization.
2. Knowledge Management (KM) comprises a range of strategies and practices used in an organization to identify, create, represent, distribute, and enable adoption of insights and experiences.
3. More recently, other fields have started contributing to KM research; these include information and media, computer science, public health, and public policy.
4. Many large companies and non-profit organizations have resources dedicated to internal KM efforts, often as a part of their business strategy, information technology, or human resource management departments (Addicott, McGivern & Ferlie 2006).
5. KM efforts overlap with organizational learning, and may be distinguished from that by a greater focus on the management of knowledge as a strategic asset and a focus on encouraging the sharing of knowledge.

Note: Available at: http://en.wikipedia.org/wiki/Knowledge_management (accessed 30 January 2011).

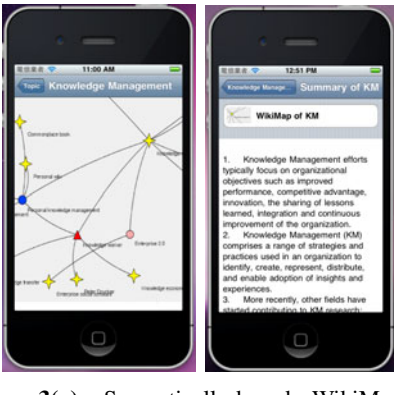

Fig. 3(a). Semantically-based WikiMap Presented in the Mobile Device

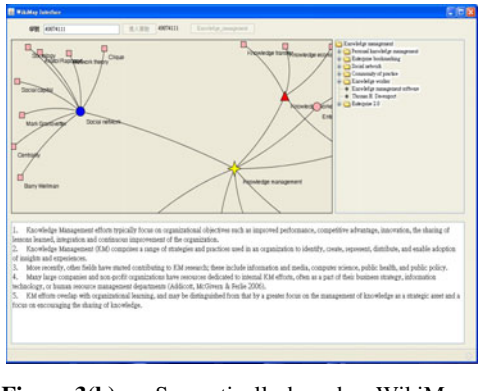

Fig. 3(b). Semantically-based WikiMap Presented in the PC

7 Conclusion and the Future Work

In this present study, we proposed an SNA-based summarization model and the development of search or navigation applications to guide users to find and organize needed information or topics within Wikipedia. We employed the ***NGD*** algorithm in the proposed ***LS*** measure to quantify the strength of the semantic relationships between articles in the topic network. Our preliminary evaluation results demonstrate the effectiveness of applying semantic analysis in an internal link-based network. To help users quickly read topically related articles, we proposed SNA-based summarization to present single article's summaries in a newly developed interface. This study also presents the interface for the semantically-based WikiMap as it would appear on PCs and mobile devices. In the future, we will evaluate the precision and accuracy of the summaries for the articles based on the proposed methods. Moreover, the effectiveness of the proposed applications will be evaluated in the real setting.

Acknowledgments. This research was supported by the National Science Council and Fu-Jen Catholic University of Taiwan under the Grant No. 99-2410-H-030-047-MY3 & No.409931074078, respectively.

References

1. Almpanidisa, G., Kotropoulos, C., Pitas, I.: Combining Text and Link Analysis for Focused Crawling—An Application for Vertical Search Engines. Information Systems 32(6), 886–908 (2007)
2. Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine. Computer Network 30(1-7), 107–117 (1998)
3. Cilibrasi, R., Vitányi, P.: The Google Similarity Distance. In: IEEE Transactions on Knowledge and Data Engineering, pp. 370–383. IEEE Press, New York (2007)
4. Evangelista, A., Kjos-Hanssen, B.: Google Distance between Words. Frontiers in Undergraduate Research. University of Connecticut (2006)
5. Finkelstein, L., Gabrilovich, Y.M., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Transactions on Information Systems (TOIS) 20(1), 406–414 (2001)
6. Forsyth, R., Rada, R.: Adding an Edge in Machine Learning: Applications in Expert Systems and Information Retrieval. Ellis Horwood Ltd., pp.198-212 (1986)
7. Hahn, U., Mani, I.: The Challenges of Automatic Summarization. Journal of IEEE Computer 33(11), 29–36 (2000)
8. Hovy, E., Lin, C.Y.: Identifying Topic by Position. In: Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP), Washington, DC, pp. 283–290 (1997)
9. Liu (ed.): Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, New York (2007)
10. Radev, D.R., Jing, H., Stys, M., Tam, D.: Centroid-based Summarization of Multiple Documents. Information Processing and Management 40(6), 919–938 (2004)
11. Teufel, S., Moens, M.: Sentence extraction as a classification task. In: Proceedings of the Workshop on Intelligent Scalable Summarization. ACL/EACL Conference, Madrid, Spain, pp. 58–65 (1999)
12. Wasserman, S., Faust, K. (eds.): Social Network Analysis: Methods and Applications. Cambridge University Press, UK (1994)
13. Wu, I.-C., Wu, C.-Y.: Using Internal Link and Social Network Analysis to Support Searches in Wikipedia: A Model and Its Evaluation. Journal of Information Science 37(2), 189–207 (2011)