Generation of Hypertext for Web-Based Learning Based on Wikification

Andrew Kwok-Fai Lui, Vanessa Sin-Chun Ng, Eddy K.M. Tsang, and Alex C.H. Ho

School of Science and Technology, The Open University of Hong Kong
{alui,scng,s1011965,s1024743}@ouhk.edu.hk

Abstract. This paper presents a preliminary study into the conversion of plain text documents into hypertext for web-based learning. The novelty of this approach is the generation of two types of hyperlinks: links to Wikipedia article for exploratory learning, and self-referencing links for elaboration and references. Hyperlink generation is based on two rounds of wikification. The first round wikifies a set of source documents so that the wikified source documents can be semantically compared to Wikipedia articles using existing link-based measure techniques. The second round of wikification then evaluates each hyperlink in the wikified source documents and checks if there is a semantically related source document for replacing the current target Wikipedia article. While preliminary evaluation of a prototype implementation seemed feasible, relatively few self-referencing links could be generated using a test set of course text.

Keywords: web-based learning, hypertext generation, wikification, wikipedia.

1 Introduction

A hypertext medium offers readers a number of advantages such as providing non-linear navigation for seeking related information, creating opportunities for exploring new knowledge, and even facilitating the generation of new knowledge. The set up of hyperlinks between documents allows the continued expansion of hypertext as exemplified by the World Wide Web. Hyperlink is made up of text anchor in the source document and the location of the target document, and the target documents of the hyperlinks are believed to be useful for explanation, reference, or further exploration. Hypertext authoring involves careful consideration of the overall objectives, the needs of readers, and the relevance of target documents.

Authoring of hypertext learning materials has meant a lot of hard work for adopters of web-based learning. The conversion of existing course notes into electronic form is the relatively easier part. The hard part is to transform the linear electronic course notes into hypertext made up of non-linear set of linked documents. The original text has to be manually inspected, segmented, reorganized, and hyperlinked. The demand to make consistent and systematic decisions about the location of hyperlink, the anchor-text, and the target of hyperlink is often too much for individual instructors. Since the early adoption days of web-based learning, automatic construction of

R. Kwan et al. (Eds.): ICT 2011, CCIS 177, pp. 280–290, 2011.

hypertexts has attracted some attention. The majority of notable previous attempts is based on text mining and information retrieval techniques.

This paper describes an approach of automatic conversion of plain course texts into hypertexts based on wikification. Wikification is a recent area of research that studies automatically enriching a piece of text with links to the online encyclopaedia Wikipedia [5][6]. The approach supports both generation of self-referencing links for self-contained study and Wikipedia links for exploratory study. Logically segmented plain texts are first wikified by adding anchors linking to specific pages in the Wikipedia. The wikified text segments are then considered as a virtual extension to the Wikipedia. The wikification process is repeated only this time the wikified text segments are also considered as potential targets of hyperlinks. The resulting hypertexts can therefore contain self-referencing hyperlinks as well as links to Wikipedia.

2 Background

Web-based learning offers a number of technological features that are relevant to educators [6]: (1) hypertext provides effective organization and ready access to vast amount of information; (2) communication medium offers opportuities for interaction and collaboration; (3) authoring tools enable everyone to create content and make available to others; (4) web-based learning environments integrate instructional activities into one delivery medium that now typically contains content management, student management, discussion forums and even weblogs. The first feature is arguably the most demanding on the effort of instructors. The development of a new hypertext based course needs a great deal of design, planning, and organization, and the level of complexity exceeds that of a traditional course. Conversion from existing course text into hypertext makes more economical sense, especially if lots of time and effort have been spent on writing the orignal course text.

2.1 Quality and Types of Hyperlinks

Hypertexts are conduicive to learning if the quality of hyerlinks are satisfactory. There should be a specific purpose for each hyperlink. Hypertext authors should ensure that the set of hyperlinks would fulfil the overall learning objectives. Instructionally, hyperlinks can be classified as structural links (ie. connect to another unit of hypertext), reference links (ie. direct to the source of the content), and associative links (leads to related concepts) [1]. The set up of associative links is more effort intensive because it requires deeper understanding of semantic relations [11]. The nature of hypertext suits the style of exploratory learning particularly well. The variety and complexity offered by hyperlinks enhances the motivation in an autonomous exploration of knowledge [9]. For examples, reference links help to elaborate anchor texts for a better understanding of a document, and associative links offer related topics to satisfy a curious mind.

2.2 Quantity of Hyperlinks in Hypertext

Hypertext based exploratory learning relies on sufficient amount of hyperlinks in a document. The so-called learning impasses describes the undesirable situation that a

lack of hyperlinks restricts opportunities to find elaborations, references, and further topics to study [9]. Such a document represents a dead-end and a hinderance to the effectiveness of exploratory learning. Clearly adding more hyperlinks to the document can resolve the problem but there is a consideration related to the cost of expertise and time. An alternative approach is to exploit the intelligence of the mass and allow learners to contribute hyperlinks. The Free-Hyperlinks environment, for example, provides learners to create and share new hyperlinks in a web 2.0 collaborative manner [10].

2.3 Automatic Generation of Hypertext

Automatically adding hyperlinks to text documents is an inexpensive approach to generate useful hypertext for web-based learning. The Dynamic Medical Handbook Project was one of the first attempts of its kind [3]. A fulltext medical handbook was converted into hypertext in four steps: (1) partioning the handbook into text units based on its intrinsic hierarchical structure; (2) extracting the first words of each text unit as anchor texts; (3) adding structural hyperlinks between hierarchically related text units; and (4) indexing sematically related text units based on an essentially a bags-of-words statistical approach. Setting up links between text units in a hypertext requires a way to estimate the semantic relation of the text content. Most subsequent work in this area recognized this bascially as a text mining process of selecting anchor texts and hyperlink targets [2][11].

2.4 Wikification

Wikification is a related research area that investigates hyperlinking existing text to relevant articles in Wikipedia. Using Wikipedia as the target of hyperlnking should give educators and learners higher confidence about the content quality. Wikipedia is more than an online collaborative encyclopeadia as perceived by general public. Medelyan et. al. [4] listed a number of other perspectives of Wikipedia, including a huge corpus, a multi-lingual thesaurus, a semi-structured database, an ontology, a scale-free small-world type of network structure. Most significant to automatic hypertext generation is perhaps its semantic richness for many text mining processes. In general a wikification process involves link detection phase and disambiguation phase, which is virtually the same involved in hypertext generation from plain text. The link detection phase decides if a term should be turned into an anchor for a hyperlink. The disambiguation phase decides the most relevant target Wikipedia article for a hyperlink anchor which may have several meanings. For a wikification system called Wikify, Mihalcea & Csomai [5] proposed the use of an atribute called link probability to idenify anchor texts, which is the probability of a term used as an anchor in Wikipedia. To disambiguate the target for a hyperlink, Wikify relied on a classifier based on a text anchor's nearby terms and their part-of-speech. The classifier is trained with examples extracted from Wikipedia with reasonably good accuracy. The promising performance of a machine learning approach prompted Milne & Witten [6] to treat link detection also as a classification problem. The features used to

predict if a term should be tuned into a hyperlink anchor include link probability, relatedness of the term to the surrounding context, generality of the term, and the location of term in the article. Using the examples in Wikipedia as the gold standard, the performance of the link detector was found to achieve 74% in precision and in recall.

3 Methodology

This section describes Hyperizer, an automatic text to hypertext converter. Text is assumed to be a set of paritioned course learning materials such as lecture notes, tutorial notes, technical manual in plain text format. Compared to earlier systems, Hyperizer is able to (1) generate links to Wikipedia articles for further explanation and exploration, and (2) generate self-referencing links for elaboration and references. The novelty of Hyperizer lies in the central role played by wikification in the generation of both types of hyperlinks.

Setting up self-referencing links between a set of text documents invariably needs a reliable way to work out the semantic relatedness between two text documents and between a term and a text document. Milne & Witten [7] developed a Wikipedia Link Based Measure that estimates the semantic similarity of two Wikipedia articles by comparing their sets of incoming or outgoing links. According to the evaluation done by Medelyan et. al. [4], the algorithm achieved a respectable 0.69 correlation on a gold standard test set and it was found to be the best among the algorithms that do not rely on deep text analysis. This algorithm can be applied on any text documents after they have been wikified. Wikification adds links to the text documents and enables the evaluation fo their semantic relatedness.

Hyperizer performs two rounds of wikification on a set of source text documents (see Fig. 1 and Fig. 2). The first round converts the text documents into hypertext documents. The added links all lead to a Wikipedia article. Hyperizer uses a machine learning approach based on the algorithm proposed by Milne & Witten to wikify the documents [6]. A full set of Wikipedia articles are needed in the training and also in the consideration of the targets of the hyperlinks. The significance of the first round of wikification is that the set of wikified documents can be regarded as Wikipedia articles for the rest of the processing. In the second round of wikification, the set of newly wikified documents has joined the existing Wikipedia articles to become potential targets for hyperlinks.

The following will describe the design of a prototype implementation of Hyperize, and will also demonstrate the conversion of a Chinese History course into a hypertext course. This version of Hyperize is designed to process Chinese text and a dump of Chinese Wikipedia is used as the corpus for training Hyperize. The techniques employed in Hyperize, however, are mostly language independent.

The wikification process of Hyperizer is based on the link detection and disambiguation algorithms proposed in [6]. The following gives a summary for each of the two algorithms. Readers may refer to the original paper for the details.

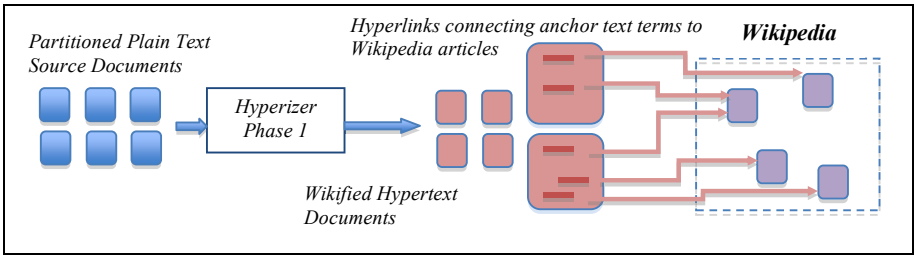

Fig. 1. Phase 1 of Hyperizer: source documents are wikified

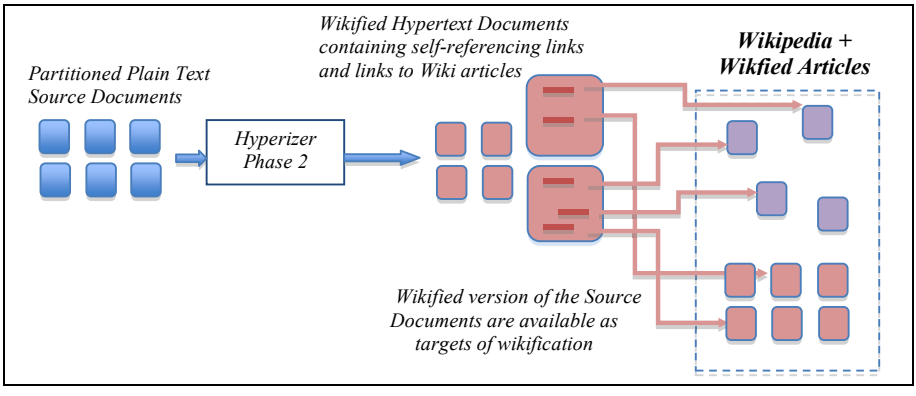

Fig. 2. Phase 2 of Hyperizer: wikified source documents after phase 1 are now considered as part of the extended Wikipedia for another round of wikification

3.1 Wikification: Disambiguation

Disambiguation algorithm is discussed first because it is part of the link detection algorithm [6]. The objective of disambiguation is to select the target Wikipedia article that matches the context of an anchor text term. For example, the term Liverpool can refer to the England city or the Premier League soccer team depending on the context of the source article. A classifer is built that evaluates the most likely target Wikipedia article. The classifier is to be trained with examples extracted from Wikipedia. There are three features used in the classifier:

- Prior probability of the link between an anchor text and a target Wikipedia article. This is obtained by mining the Wikipedia. For example, in the Chinese Wikipedia version used in the experiment, 21% of anchor text '利物浦' (Liverpool) is linked to the city's article and 79% is linked to the soccer team's article.
- Relatedness of the anchor text to the theme of the source document. This is estimated by comparing the semantic relatedness of the anchor text and the other anchor texts in the source article. The comparison is based on the Link Based Measure described in [7].
- Quality of the Relatedness feature. The reliability of the Relatedness feature depends on whether there is a central theme in the source document. This feature

estimates the cohesiveness of the theme in the source document by mutually comparing the semantic relatedness of all anchor texts.

3.2 Wikification: Link Detection

The objective of link detection is to identify the text terms for conversion into hyperlink anchors. A characteristic in the algorithm of Milne & Witten [6] is to consider only the text terms that have been used as anchor text in Wikipedia. Mining the Wikipedia can extract a table of anchor text terms. This characteristic is important for wikifying Chinese text because the problem of word segmentation can be circumvented. Chinese text has no natural boundary and most text processing operations begin with word segmentation that split up the text into terms. For example, this Chinese string "中國中古史是從秦、漢這兩個統一王朝開始的" (The ancient history of China began from the period of the unified dynasties of Qin and Han) is to be split up into terms such as "中國 中古史 是 從 秦、漢 這 兩個 統一 王朝 開始 的". In the link detection algorithm, word segmentation, which is not always highly reliable, is not needed. Instead, the table of anchor text terms is used to extract all occurrences of anchor text terms in the source document, and each of these is a potential hyperlink anchor.

Similar to the disambiguation algorithm, a classifier is used to determine if an occurrence of an anchor text term should be converted to hyperlink anchor. The classifier is to be trained with examples extracted from Wikipedia. There are eight features used in this classifier:

- Probability of the anchor text actually used as a hyperlink anchor.
- Relatedness of the anchor text to the theme of the source document. This feature is the same as the one used in the disambiguation algorithm.
- Confidence of the disambiguation classifier when applied to this anchor text.
- Generality of the anchor text mined from the category information in Wikipedia.
- Frequency of the anchor text in the source document.
- The Location of the First Occurrence, the Last Occurrence, and the Distribution of the anchor text in the source document.

3.4 Hyperizer: Conversion to Self-reference Links

The wikified documents *Ws1, Ws2, …, Wsn* generated by the first phase of Hyperizer contain new hyperlinks to Wikipedia articles. They can be semantically compared to Wikipedia articles and each other. The second phase of Hyperizer then redirects some of the new hyperlinks from a Wikipedia article to one of the wikified source documents.

The algorithm is described below:

1. For each wikified source document Wsi, consider every new hyperlink connecting to a Wikipedia article.

1.1 For a hyperlink Linkj in source document Wsi connecting to a Wikipedia article WKj

1.1.1 Evaluate the semantic relatedness of the article WKj with every wikified source document Ws1, Ws2, …, Wsn. Find out the wikified source document Wshigh with the highest semantic relatedness.

1.1.2 If the highest semantic relatedness is greater than a threshold SemT, change the target of the hyperlink Linkj to the wikified source document Wshigh.

The algorithm considers only the hyperlinks generated by the first phase for efficiency reason.

4 Results

A prototype system of Hyperizer has been implemented for the evaluation of the algorithms. The system is implemented with Java and WEKA, the open-source machine-learning library (http://www.cs.waikato.ac.nz/ml/weka/).

4.1 Data Set

An article dump of Chinese Wikipedia released on 25 December 2009 has been downloaded (http://dumps.wikimedia.org/). The dump contains 593,003 articles, of which 303,341 are proper articles, 231,458 are re-direction pages, and 57,215 are category pages. Efficient handling of such a gigantic structured corpus requires processing tools, such as the Wikipedia API [12]. The pre-processing stage extracts sets of useful information from the Wikipedia dump, such frequency of anchor text terms, hyperlinks of every article, categories of article, etc. These extracted information sets are placed in a database for efficient query and access.

4.2 Wikification: Disambiguation

An experiment was carried out to investigate the performance of our implementation of the disambiguation algorithm trained with the Chinese Wikipedia dump. The training set contains 1,000 randomly chosen articles and the test set contains another 1,000 random articles. All articles selected have length between 1,000 to 2,000 Chinese characters. Hyperlinks found in lists are removed because they are less relevant to the aim of Hyperizer.

Table 1 below shows the performance of our implementation. The classification algorithm used was the C4.5 decision tree algorithm. The results showed are comparable with the performance reported in [6]. Chinese anchor text probably has fewer senses and the disambiguation classifier should find it less challenging. In the calculation of the feature Relatedness, Milne & Witten [6] used at most 30 anchors to represent the theme of the document. An experiment was carried out and found that relaxing the limit would not improve the performance.

4.3 Wikification: Link Detection

Another experiment was carried out to evaluate the performance of the link detection algorithm trained with the Chinese Wikipedia dump. Link detection is considerably more challenging than disambiguation. Deciding the creation of a hyperlink requires

more considerations than choosing a target Wikipedia article between a few possibilities. The training set contains 100,000 examples and the testing set contains another 100,000 examples selected from Wikipedia articles.

Table 1. Performance of our implementation of the disambiguation algorithm

	Precision	*Recall*	*F-Measure*
C4.5 (Chinese Wikipedia)	98.3%	98.3%	98.3%
C4.5 (Chinese Wikipedia) limited to 30 anchors in calculating Relatedness	*98.3%*	*98.3%*	*98.3%*
C4.5 (English Wikipedia) disambiguation algorithm (Milne & Witten 2008a)	*96.8%*	*96.5%*	*96.6%*

Table 2 below shows the performance of our implementation. Using the Chinese Wikipedia as the corpus seems to produce poorer precision. Changing the classification algorithm to support vector machine gave a bit of improvement. The overall performance is comparable to Milne & Witten in [6].

Table 2. Performance of our implementation of the link detection algorithm

	Precision	*Recall*	*F-Measure*
C4.5 (Chinese Wikipedia)	71.2%	78.0%	74.4%
Support Vector Machine (Chinese Wikipedia)	73.7%	74.3%	74.0%
C4.5 (English Wikipedia)	77.6%	72.2%	74.8%
link detection algorithm (Milne & Witten 2008a)			

4.4 Hyperizer: Conversion to Self-reference Links

A set of source plain text documents was prepared for this experiment. The document set comes from a course in Chinese History. This theme is one of the most popular topics in Chinese Wikipedia and so wikification should produce an interesting lot of hyperlinks. The document set was first manually partitioned into thirty-six documents, each of size from around two thousand words to over five thousand words. The titles of some of the partitioned documents include 秦興起及統一過程 (The Rise and Unification of Qin), 秦始皇的統治政策 (The Rule and Policy of Emperor Qinshihuang), 秦朝的覆亡 (The Fall of Qin Dynasty), 秦亡原因 (Reasons of the Fall of Qin), and 漢初對秦政的因革 (Reform of Qin Ruling Style in Early Han). This granularity is comparable to a typical overview article in the Wikipedia. For example, the article on "秦朝" (Chin Dynasty) has approximately 4,600 words.

The following illustrates the operation of Hyperizer with the test document set. Fig 3 (top) shows a segment of text from the document 秦亡原因 (Reasons of the Fall of Qin) and the bottom shows all the anchor text terms (underlined) found in the segment.

秦始皇在政治上所作的改革，無疑是劃時代的，應該肯定他在這方面的功績。可是秦始皇在政治上的極權表現，卻削減了在當時本來帶有進步意義的政制改革的作用。在「丞相、大臣皆受成事，倚辦於上」，「天下之事無小大，皆決於上」（《史記・秦始皇本紀》）的政治情況下，任何完善的政治規劃，都不可能確保發揮其應有的效用。

秦始皇在政治上所作的改革，無疑是劃時代的，應該肯定他在這方面的功績。可是秦始皇在政治上的極權表現，卻削減了在當時本來帶有進步意義的政制改革的作用。在「丞相、大臣皆受成事，倚辦於上」，「天下之事無小大，皆決於上」（《史記・秦始皇本紀》）的政治情況下，任何完善的政治 規劃，都不可能確保發揮其應有的效用。

Fig. 3. An example text segment and the anchor text terms found

After the first round of wikification, only some of the anchor text terms were converted into hyperlinks (see Fig 4). Table 3 lists the target Wikipedia article of each hyperlink.

秦始皇在政治上所作的改革，無疑是劃時代的，應該肯定他在這方面的功績。可是秦始皇在政治上的極權表現，卻削減了在當時本來帶有進步意義的政制改革的作用。在「丞相、大臣皆受成事，倚辦於上」，「天下之事無小大，皆決於上」（《史記・秦始皇本紀》）的政治情況下，任何完善的政治規劃，都不可能確保發揮其應有的效用。

Fig. 4. The wikified version of the text segment

Table 3. Hyperlinks and their target Wikipedia article generated by first round wikification

Anchor Text Term	*Target Wikipedia Article (Chinese)*	*Corresponding Article in English Wikipedia*
秦始皇（QinShiHuang)	秦始皇 (QinShiHuang)	Qin Shi Huang
丞相 (Premier)	宰相 (Premier)	Chancellor
史記 (Ancient History)	史記 (Ancient History)	Records of the Grand Historian
秦(Qin)	秦朝 (Qin Dynasty)	Qin Dynasty

秦始皇在政治上所作的改革，無疑是劃時代的，應該肯定他在這方面的功績。可是秦始皇在政治上的極權表現，卻削減了在當時本來帶有進步意義的政制改革的作用。在「丞相、大臣皆受成事，倚辦於上」，「天下之事無小大，皆決於上」（《史記・秦始皇本紀》）的政治情況下，任何完善的政治規劃，都不可能確保發揮其應有的效用。

Fig. 5. A self-reference link has replaced a hyperlink to Wikipedia (highlighted in HTML)

In the second round of wikification, the target Wikipedia page of each hyperlink was used to semantically compare to all the thirty-six wikified source documents. The source document with the highest link measure and over the threshold was chosen as

a self-reference link. Only one in four hyperlinks was replaced (see Fig. 5). In fact, this was the only self-referencing link converted by Hyperizer in the whole document. While there was a successful conversion, the conversion rate was however rather disappointing.

5 Conclusion

This paper reports a preliminary study into the topic of the conversion of a set of plain text documents into hypertext for web-based learning. A prototype implementation called Hyperizer was design for the purpose, and it was developed as a proof of concept. Unlike earlier work in this problem, Hyperizer can generate links to both Wikipedia articles and within the original text documents. The former type of links enhances exploratory learning while the latter type provides references and elaboration to learners.

The key idea of Hyperizer is to first wikify the set of source documents, so that these wikified source documents can be semantically compared to Wikipedia articles using current semantic comparison techniques.

The evaluation showed that the wikification algorithms developed by Milne & Witten [6] performed equally well when they were using Chinese Wikipedia as the training corpus. The link detection algorithm scored slightly lower precision on Chinese Wikipedia but the overall performance was still comparable.

Finally, the experiment on Hyperizer illustrated that the feasibility of the approach seemed positive. However, the conversion rate from Wikipedia hyperlink to self-referencing link looked very low. One possible reason is that only hyperlinks were considered in the semantic comparison between the wikified documents and Wikipedia articles. There are often too few hyperlinks generated by the first round of wikification. A solution to evaluate in future work would be to consider all anchor text terms. The target Wikipedia articles of these anchor text terms could be estimated with the disambiguation algorithm.

References

1. Agosti, M., Crestani, F., Melucci, M.: On the use of information retrieval techniques for the automatic construction of hypertext. Information Processing and Management 33, 133–144 (1997)
2. Crestani, F., Melucci, M.: Automatic construction of hypertexts for self-referencing: the Hyper-Textbook project. Information Systems 28, 769–790 (2003)
3. Frisse, M.F.: Searching for Information in a Medical Handbook. Communications of the ACM 31(7), 880–886 (1988)
4. Medelyan, O., Milnea, D., Legga, C., Witten, I.H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67(9), 716–754 (2009)
5. Mihalcea, R., Csomai, A.: Wikify! Linking Documents to Encyclopedic Knowledge. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-8, vol. 8, pp. 233–241 (2007)
6. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WIKIAI 2008), Chicago, IL (2008)

7. Milne, D., Witten, I.H.: Learning to Link with Wikipedia. In: Proceedings of CIKM, pp. 509–518. ACM, New York (2008)
8. Mioduser, D., Nachmias, R., Lahav, O., Oren, A.: Web-based learning environments: current pedagogical and technological state. Journal of Research on Computing in Education 33(1), 55–76 (2000)
9. Mitsuhara, H., Ochi, Y., Kanenishi, K., Yano, Y.: Adaptive Web-based Learning System with Free-hyperlink Environment for Circumventing Exploration Impasse Caused by Hyperlink Shortage. The Journal of Information and Systems in Education 1(1), 109–118 (2002)
10. Mitsuhara, H., Ochi, Y., Kanenishi, K., Yano, Y.: An Adaptive Web-based Learning System with a Free-Hyperlink Environment. In: Proc. of Workshop on Adaptive Systems for Web-based Education, pp. 13–26 (2002)
11. Yang, H.C., Lee, C.H.: A text mining approach for automatic construction of hypertexts. Expert Systems with Applications 29, 723–734 (2005)
12. Zesch, T., Gurevych, I., Mühlhäuser, M.: Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In: Data Structures for Linguistic Resources and Applications, pp. 197–205 (2007)