EASY EDUCATIONAL LITERATURE COMPREHENSION
Applying TermPedia to Information Retrieval Knowledge Domain

Proscovia Olango, Gosse Bouma
The Center for Language and Cognition Groningen, The University of Groningen, Oude Kijk in 't Jatstraat 26, Groningen, The Netherlands
p.olango@rug.nl, g.bouma@rug.nl

Henny Klein
Department of Information Science, The University of Groningen, Groningen, The Netherlands
E.H.Klein@rug.nl

Keywords: Document enrichment, term extraction, term definition, term disambiguation, hyperlink generation, document comprehension.

Abstract: This paper provides a detailed description of how TermPedia which is a document enrichment tool, is applied to educational literature in order to help students easily understand their reading material. The paper assumes that technical terms are one of the major hindrances to document content comprehension. TermPedia has the ability to extract, define, and link technical terms to Wikipedia. It is expected that relevant term definitions and explanations will ease the comprehension of educational documents by students and thereby improve their reading speed and shorten the time they needed for knowledge acquisition.

1 INTRODUCTION

Easy document comprehension can be defined as the ability to read and understand the concepts discussed in a document without spending much time and brain effort. In cognitive science, comprehension is often characterised as the construction of a mental model that represents the objects and semantic relations described in a text (Thuring et al., 1995). This implies that if a reader spends less effort in the construction process, then the document is easily comprehensible.

The field of document comprehension is widely studied with reference to reading levels indicating that persons who are not reading at a college level or higher, have a low reading level and therefore do not clearly understand the concepts of the documents they read (Young et al., 1990). These studies indicate that education plays an important role in vocabulary level building which in turn eases understanding of semantically related documents.

As an implementation of research findings based on reading levels, vocabulary (or technical terms) that occur in documents are substituted with simpler words or word phrases that have similar semantics (Graves and Graves, 2003). If the substitution is done without considering word context, this venture may not only distort document content meaning but also limit the possibility for students to build their vocabulary level. By technical term we refer to anything that is a word, a group of words, an acronym, or an abbreviation which designates a special meaning in text context. Research in reading shows substantial evidence that the meaning of difficult vocabulary can be derived from context (Kate, 2007), a concept that has been largely applied by computational linguists.

However, human languages more especially English constantly change by borrowing, coining, and combining words to represent new ideas, [technology,] and development (Engineer, 2005). Therefore vocabulary level building becomes a lifetime obligation.

> In 2009 the verb **twitter** was borrowed as trade mark of a social network that provides **microblogging** services, enabling its users to send and read messages called **tweets**.

The example above presents three technical terms that were borrowed or coined in the recent years to express ideas related to Short Message Services (SMS) and Internet technology. For a person who is not familiar with these technologies the semantics of the terms *twitter*, and *tweet* may be conspicuously re-

vealed by considering them in context. On the contrary the meaning of the term *microblogging* may not be that obvious from this sentence context.

Technical terms occur in almost all reading material especially those intended for an audience at higher institutions of learning like universities and they may hinder content comprehension if they are introduced without definitions and/or explanations. We assume that easy access to context related definitions and explanations of technical terms found in a document will simplify document comprehension for readers of all levels without depriving them the opportunity to build their vocabulary levels. Therefore, TermPedia was designed with an objective to provide easy access to contextually relevant definition of technical terms that are embedded in documents. In case a technical term definition is obscure, TermPedia contains an option for the reader to obtain additional explanation on the defined technical term by linking to an article that discusses it explicitly.

1.1 TermPedia Technologies

TermPedia is a document enrichment tool that uses Human Language Technologies (HLTs) to provide contextually relevant information for technical terms that are embedded in documents. TermPedia was designed to define technical terms by extracting their meaning from an on-line encyclopedia called Wikipedia. Wikipedia is a multilingual web-based free-content encyclopedia project based on an open-editable model. Although Wikipedia offers multilingual documents, we only utilize the English [1] content.

Defined technical terms are linked to contextually relevant Wikipedia articles so that in cases where a term definition does not provide sufficient information for content comprehension, a reader may navigate to the Wikipedia article for additional explanation. The HLTs used in TermPedia include semi-automatic technical term extraction, automatic term definition and automatic hyperlink generation.

Much as document enrichment [that is to say, incorporation of contextually relevant information into existing text] is the fundamental technology of TermPedia, the tool uses document extraction techniques for semi-automatic term extraction. In addition a simple string look-up algorithm is used for automatic term definition and a frequency based algorithm is used to generate automatic hyperlinks (Olango et al., 2009).

[1] See http://en.wikipedia.org for the English Wikipedia

1.1.1 Semi-Automatic Term Extraction

The term extraction process of TermPedia is semi-automatic because we use a pre-defined list of terms. The list is created by extracting Wikipedia anchor texts, where each anchor text is treated as a technical term. An anchor text is the alternative set of characters that are displayed in place of a web address after the creation of a hyperlink. For example, *twitter* is an anchor text for the hyperlink below.

```
<a href=``http://en.wikipedia.org/wiki/
Twitter''>twitter</a>.
```

Major links in Wikipedia include external, internal, and related links. For TermPedia, anchor texts were extracted from the later two types of links that reference Wikipedia articles and articles from Wikipedia sister projects like Wiktionary. The advantage of using these anchor texts is that they provide many alternative phrases all linked to the same concept. Thus, linking technical terminology in an educational document becomes easier, since the probability that the same phrase has been used as an anchor text is considerably high given the coverage depth of Wikipedia.

1.1.2 Automatic Term Definition

A close look at Wikipedia articles reveals that the first paragraph normally provides a detailed definition of the title. This information is used to provide definitions for the terms that where predicted by help of Wikipedia anchor text. Assuming that each predicted term represents a Wikipedia article title, then the first paragraph of an article whose title string matches the predicted term is taken as the definition of that particular term.

The current definition of a term is extracted from Wikipedia using a Hypertext Pre-Processor (PHP) script. PHP is a scripting language designed for web development to produce dynamic web pages and is used here to retrieve a Wikipedia article that is relevant to the term of interest. The PHP script embeds a Perl script which is responsible for matching terms to Wikipedia article titles using a string look-up algorithm. If the term definition is insufficient for content comprehension, a user may navigate to a Wikipedia article for additional explanation on that term through a hyperlink that is automatically generated for each extracted term.

1.1.3 Automatic Hyperlink Generation

A hyperlink is a reference from one electronic document to another that can be triggered by a user usually through computer mouse actions like clicking and

hovering. In TermPedia, hyperlinks are automatically generated by using the predicted terms as anchor text for an HTML (HyperText Markup Language) $< a >$ tag using a frequency-based term extraction (FTE) algorithm. FTE is based on the keyword ranking method called Keyphraseness, which was presented by (Mihalcea and Csomai, 2007). In this approach all possible n-grams in a document that are present in the list of terms are identified and ranked according to their likelihood of being selected as a technical term (TT). If a term is most of the time selected as a TT among its total number of occurrence, it is most likely that it will again be selected in a new document as a TT. Therefore the probability (*P*) that a term *(X)* is selected as a TT in a new document is calculated as the total number of documents where the term was already selected as a TT *(count(D_{TT}))* divided by the total number of documents where the term appeared *(count(D_X))*.

$$P(TT|X) \approx \frac{count(D_{TT})}{count(D_X)} \quad (1)$$

We select only the top 7% (according to keyphraseness) of technical terms in a document, thus preventing the generation of hyperlinks for common words and/or avoiding saturating the text with irrelevant links. The *href* attribute value of the $< a >$ tag is indicated by the web address of the Wikipedia article to which the anchor text originally referred.

```
Introduction to <a href =
''http://en.wikipedia.org/wiki/
Information_Retrieval''>Information
Retrieval</a>.
```

In the above example, *Information Retrieval* is the extracted term and anchor text for the $< a >$ tag and $http://en.wikipedia.org/wiki/Information\_Retrieval$ is the value of the *href* attribute. So far, we have discussed how TermPedia extracts terms and generates hyperlinks section 3.3 discusses a web-based user interface that was designed to extend TermPedia to university students.

2 RELATED WORKS

Work that has been done in relation to document enrichment and eLearning, term and definition extraction, and automatic hyper-link generation, are discussed in this section with a focus on their importance to improving educational literature comprehension.

2.1 Document Enrichment and eLearning

With an intention to simplify document comprehension by artificial extension of a reader's knowledge-base on a per-need basis (Csomai and Mihalcea, 2007), presented Wikify! a system that can improve educational materials by automatically extracting keywords, technical terms and other key concepts and linking them to appropriate Wikipedia articles (a concept they referred to as "text wikification"). After carrying out a test by selecting 14 quiz questions from an on-line history course taught at the University of North Texas, the test results showed that bringing information relevant to the topic of student's study within their easy reach through hyperlinking is a successful strategy for increased effectiveness in pedagogical tasks. Technically, Wikify! is a document enrichment tool that has shown basic success in improving educational literature comprehension by use of technical terms. Nowhere though does Wikify! consider automatic term definition which TermPedia proposes as a key for improving educational literature comprehension. The time required to acquire knowledge may be reduced since a student does not need to link to an external source if the term definition provides adequate information for comprehension.

(Monachesi and Westerhout, 2008) Worked on a Language Technology for eLearning (LT4eL) project where they adapted results from Natural Language Processing (NLP) to eLearning context by using statistical measures in combination with linguistic processing to detect keyword candidates. The key words where then used as glossary candidates which allowed for the creation of glossaries based on definition of the relevant terms, to be linked to learning objects. Whereas these authors focused on how to find definitions of terms, we concentrate on integrating the links to definitions, and on context-dependent disambiguation to provide relevant information for content comprehension.

2.2 Term Extraction and Definition

Linguistic patterns, statistical methods, tf-idf (term frequency-inverse document frequency) weight, term co-occurrence, and concept identification using Wikipedia are methods that have been used in term extraction (Fahmi et al., 2007; Medelyan et al., 2008; Aguilar et al., 2010). The latter considers Wikipedia page articles as terms and these are in turn used to extract terms in other documents. (Medelyan et al., 2008) reported that the Wikipeida technique was significantly more effective than the other tech-

niques, this report reinforced our motivation for using Wikipedia titles and links for term detection. The report also mentions that tasks such as word sense disambiguation and word similarity could be automatically addressed by exploiting Wikipedia's unique features. We however, have exploited a unique feature of Wikipedia to perform automatic term definition as explained in section 1.1.2 above.

2.3 Automatic Hyperlink Generation

Relative effectiveness of link generation based on Wikipedia article names or titles was investigated by (Fachry et al., 2008). After experimenting with Vector Space Model (VSM) and sub-string match for detecting missing links in Wikipeida, they showed that exact sub-string matching indicates an improvement in finding the missing links. This investigation performs a filtering which considers sub-strings that only match Wikipedia article title but, we consider all sub-strings collected from previous Wikipedia links as potential candidates for link generation in educational literature. In addition to link generation, we also provide relevant definitions for the sub-strings used in the link creation with and objective of reducing the time for information search and improving document comprehension.

3 APPLICATION OF TERMPEDIA

This section discusses the application of TermPedia at the University of Groningen in the Netherlands with the prospects of surveying the tool's usefulness. It is worth mentioning that the user survey was in progress at the time this paper was submitted to the conference for review. In the mean time, the survey was completed and a brief discussion of the results is presented in section 4. Students of information retrieval were selected as users for this survey because TermPedia is a practical application of information retrieval and also because the textbook for this course is available in soft copy[2].

3.1 Creation of Information Retrival (IR) Data for Annotation

We exported Wikipeida pages wrapped in XML (EXtensible Markup Language) in the category of information retrieval and other related fields like computer science and computational linguistics by using the "Export pages" feature of Wikipedia. The exported pages were stored in a file of 135 megabytes large and contained 18,685 articles. By extracting Wikipedia internal, and related links from these articles, we were able to obtain approximately 166,974 terms which in turn referred to 159,507 Wikipedia articles. This shows that each of the exported Wikipeida articles contained at least 9 links and each link could possibly refer to more than one Wikipedia page. Since the links refer to more than one Wikipeida page this shows that the links are ambiguous and should be disambiguated in reference to text context while using them for generating annotations in the information retrieval textbook.

Table 1 below shows the possible ambiguity in the term *tables*. We see that this term could refer to a *mathematical table* or a *database table*. Which meaning is relevant can only be determined in relation to text context and the TermPedia feature of automatic link generation discussed in section 1.1.3 helps to reduce this ambiguity. From the table we also see that the abbreviation *IR* that refers to *Information Retrieval* is taken care of as an advantage of using Wikipedia links to represent terms. Notice that all possible definitions of each technical term is extracted along with the Wikipeida article to which that term refers. All this information is stored in a database which makes TermPedia a relatively first tool for extraction and definition of technical terms and automatic hyperlink generation.

Table 1: Random selection of Terms and name of Wikipeida articles to which they link.

Terms	Title of Wikipeida Articles	Definitions
IR	IR	Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web...
	information retrieval	-do-
tables	Mathematical table	Before calculators were cheap and plentiful, people would use mathematical tables -lists of numbers showing the results of calculation with varying arguments? to simplify and drastically speed up computation...
	mathematical table	-do-
	Table (database)	In relational databases and flat file databases, a table is a set of data elements (values) that is organized using a model of vertical columns (which are identified by their name) and horizontal rows...
	database table	-do-

3.2 Annotation of IR Book

As mentioned in section 1.1, a string look-up algorithm is used to extract technical terms embedded in the information retrieval text book. The string look-up algorithm matches sub-strings from the database

[2]See, *http://nlp.stanford.edu/IR-book/information-retrieval-book.html*

that contains technical terms to sub-strings from the information retrieval textbook. Each strict match in the textbook is then annotated as a technical term and provided with a definition. The annotation is then transformed into an automatic html hyperlink by using the matched sub-string as an *anchor text* and providing a contextually relevant Wikipedia article as a value for the *href* attribute. Manual correction was done to make sure that each annotated technical term (matched sub-string) referred to a contextually relevant Wikipeida article thereby providing a relevant definition for that term.

3.3 Integration of Annotated IR Book into Electronic Content

In order to present TermPedia to the information retrieval students we designed a web-based user interface that integrated the annotated book into Nestor. Nestor is the digital learning environment of the University of Groningen. A screen shot of the demonstration web-based user interface is given in figure 1.

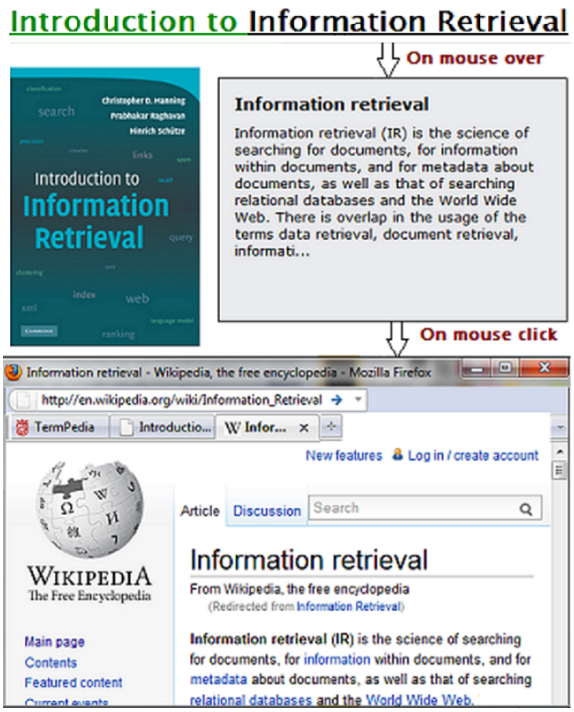

Figure 1: Screen Sorts of TermPedia User Interface

The user interface[3] allows a student to read an annotated HTML version of the information retrieval textbook with specific emphasis on chapters six to nine. Each automatically created hyperlink has an added [javascript] functionality, that allows definitions of the link (i.e. the first paragraph of the corresponding wikipedia page) to show in a pop-up window, as soon as a students moves a mouse over the link (mouse hovering action). Definitions are retrieved in real-time from the current version of Wikipedia with an excellent speed. In addition, all existing generated hyperlinks on the page are turned into links which point to the system, so that any new page accessed by the student is also automatically enriched.

[3]On-line at, *http://www.let.rug.nl/ olango/TermPedia/*

4 TERMPEDIA USER SURVEY

After the students had used TermPedia for one month they answered a quantitative questionnaire that investigated the usefulness of the tool. This part of the user survey constituted a manual method of data collection and analysis. An automatic method of data collection and analysis was also done using Google Analytics, a free web analytics tool offering detailed visitor statistics (MAS11, 2011). Just before introducing TermPedia to the students, they were given terminology exercises in order to prepare them for the possible advantages of the tool in identifying, defining and explaining difficult terms.

4.1 Manual Data Collection and Analysis for TermPedia User Survey

As mentioned above, a quantitative questionnaire was used in the manual collection of data during the TermPedia user survey. The questionnaire was divided into three sections that surveyed difficulty of the information retrieval textbook, the possibility that TermPedia is a useful tool for easy educational material comprehension and design of TermPedia user interface. The questionnaire demanded ranked responses from 0 to 4 as indicated in table 2.

Table 2: Ranks of answers to survey questionnaire.

Strongly agree	Agree	Somewhat agree	Disagree	Strongly disagree
4	3	2	1	0

Questions in the textbook difficulty section concentrated on whether the students found technical terms that they could not easily understand without the help of additional explanation and if these terms hindered there understanding of the textbook content.

The section that surveyed the possibility that TermPedia is a useful tool for easy educational literature comprehension included questions on whether the tool was able to identify technical terms from the textbook correctly and accurately provide their definitions and external link to a relevant Wikipedia article. The section also included questions on whether the students were able to easily understand the content of the textbook after accessing the definition of technical terms or rather if they needed additional explanation of the terms in addition to their definition. The last section of the questionnaire on TermPedia user interface asked questions on whether it was easy to find information on the user interface, easily navigate through the interface and also if the font sizes, colours, and formates where legible.

Five (5) questions were selected from each section for analysis of the survey results. We receive 8 responses out of 13 students who attended the Information Retrieval class. Cronbach's alpha relability coefficient was used to determine if the users' ranks were internally consistent. Alpha was calculated using the following formula:

$$\alpha = \frac{rk}{[1+(k-1)r]} \tag{2}$$

Where: k = number of items considered
and r = mean of the inter-items correlations

Cronbach's alpha reliability coefficient normally ranges between 0 and 1. The closer it is to 1 the greater the internal consistency of the items in the scale (Gliem and Gliem, 2003). Hypothesis for Cronbach's alpha relability coefficient were set as follows:

- H_1 $\alpha \neq 0$ i.e. Internal consistency of user ranks in this scale
- H_0 $\alpha = 0$ i.e. No internal consistency of user ranks in this scale

Table 3: Cronbach's alpha reliability coefficient

Survey Section	**Items**	**Mean Scores**	α	**Interpretation**
Book difficulty	5	2.4	0.3	Poor / unacceptable
TermPedia usefulness	5	2.7	0.6	Questionable
TermPedia user interface	5	2.5	0.7	Acceptable

Table 3 indicates alpha values for the three sections of the quantitative questionnaire along with their mean ranks as answered by 8 students following the Information Retrieval class. Although the mean rank of the Information Retrieval book difficulty section is slightly above average (i.e. $2.4 > 2.0$), the students' ranks for this section are inconsistency since $\alpha = 0.3$. It can therefore be said that the level to which each student found the information retrieval book difficult was subjective to that particular student. An explanation for the students' inconsistency in interpreting the book difficulty could be that they are at different vocabulary levels of Information Science knowledge domain. Since a total of 8 respondents is rather small, it can be said that $\alpha = 0.6$ is acceptable for internal consistency of students ranks for questions in the section of TermPedia usefulness. The mean rank of 2.7 for this section shows that the students consistently agree that the tool is useful for easy comprehension of educational material. The web-based user interface section gave the best results for α with $\alpha = 0.7$ and rank mean of 2.5 indicating that the students unanimously agreed that it was easy to access information from and navigate through TermPedia user interface.

4.2 Automatic Data Collection and Analysis for TermPedia User Survey

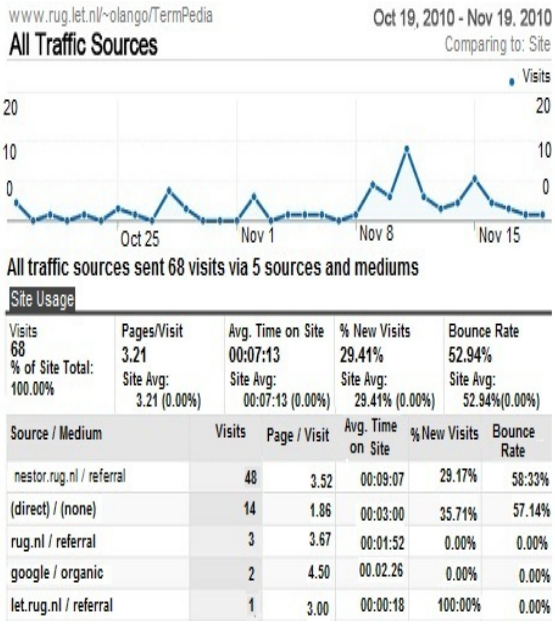

Figure 2: Statistics for TermPedia Site Usage

The automatic collection and analysis of TermPedia survey data was done using Google Analytics to provide proof that the students actually used TermPedia, therefore their responses to the quantitative questionnaire could be trusted. The advantage of using Google

Analytics is that detailed visitor statistics are automatically generated. From figure 2 we see that there were a total of 68 visitors during the one month of TermPedia usage by Information Retrieval students. The figure also reveals that 48 visitors out of the 68 (i.e. 70.6%) of the visits to TermPedia web-based user interface was from *nestor.rug.nl*, the server on which TermPedia was integrated into the students electronic content. Thus we can confidently say that the students used TermPedia with a pick of 10 visit a day during the month of November, 2010.

5 CONCLUSIONS

We have shown that TermPedia can be successfully integrated into electronic content of educational literature. By use of TermPedia students can easily access the definition of technical terms that occur in educational literature and link to Wikipedia for further explanation on the terms if the definitions do not provide adequate information for content comprehension. Since TermPedia takes care of term ambiguity, students are brought closer to accurate information, this is expected to reduce the time required for knowledge acquisition and understanding. The automatic hyperlinking feature of Termpedia also reduces the cost of maintaining unstructured information.

In Future, an experiment will be carried out to verify if TermPedia actually reduces the time for knowledge acquisition. In addition, frequency counts of activated links to Wikipedia articles whose anchor texts are the extracted technical terms contained in TermPedia annotated text shall reveal a pattern in the terms that students often look-up. It is possible that the pattern in user preferences shall expose technical terms that the students find difficult to understand. This information shall also highlight which technical terms could not be understood by the students through their definitions.

REFERENCES

Aguilar, C., Acosta, O., and Sierra, G. (2010). Recognition and extraction of definitional contexts in spanish for sketching a lexical network. In *YIWCALA '10: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 109–116, Morristown, NJ, USA. Association for Computational Linguistics.

Csomai, A. and Mihalcea, R. (2007). Linking educational materials to encyclopedic knowledge. In *Proceeding of the 2007 conference on Artificial Intelligence in Education*, pages 557–559, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Engineer, S. (2005). 21st century english vocabulary.

Fachry, K. N., Kamps, J., Koolen, M., and Zhang, J. (2008). Using and detecting links in Wikipedia. In Fuhr, N., Lalmas, M., Trotman, A., and Kamps, J., editors, *Focused access to XML documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007)*, Lecture Notes in Computer Science. Springer Verlag, Heidelberg.

Fahmi, I., Bouma, G., and der Plas, L. V. (2007). Learning to identify definitions using syntactic features. In *EACL 2006 Workshop on Learning Structured Information in Natural Language Application*, pages 64–71.

Gliem, J. A. and Gliem, R. R. (2003). Calculating, interpreting, and reporting cronbachs alpha reliability coefficient for likert-type scales. Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.

Graves, M. F. and Graves, B. B. (2003). *Scaffolding Reading Experiences: Designs for Students Success*. Christopher-Gordon Publishers, Inc., 2nd edition.

Kate, C. (2007). Deriving word meanings from context:. *Journal of Research in Reading*, Vol.30:347–359.

MAS11 (2011). Seo and internet marketing services. Internet Marketing and SEO Glossary. [online] `http://www.midatlanticseo.com/information/internet-marketing-glossary/SEO-keywords-and-definitions-a.php`.

Medelyan, O., Legg, C., Milne, D. N., and Witten, I. H. (2008). Mining meaning from wikipedia. *CoRR*, Vol. abs/0809.4530.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 233–242, New York, USA. ACM.

Monachesi, P. and Westerhout, E. (2008). What can `NLP` techniques do for e`L`earning? In *INFOS2008 proceedings*, pages 150–156. Faculty of Computer & Information-Cairo University. Cariro, Egypt.

Olango, P., Kramer, G., and Bouma, G. (2009). Termpedia for interactive document enrichment: using technical terms to provide relevant contextual information. In *IMCSIT '09: International Multiconference on Computer Science and Information Technology*, pages 265–272. IEEE.

Thuring, M., Hannemann, J., and Haake, J. M. (1995). Hypermedia and cognition:. In *Design for Comprehension*. Community of the ACM.

Young, D. R., Hooker, D. T., and Freeberg, F. E. (1990). Information consent documents: Increasing comprehension by reducing reading level. *The Hastings Center*, Vol. 12(No.3).