Author(s): Lahti, Lauri

Title: Guided generation of pedagogical concept maps from the Wikipedia

Year: 2009

Version: Post print

Please cite the original version:

Lahti, Lauri. 2009. Guided generation of pedagogical concept maps from the Wikipedia. World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education (E-Learn 2009). P. 1741-1750. ISBN 1-880094-76-2 (printed).

Lahti, L. (2009b). Guided generation of pedagogical concept maps from the Wikipedia. Proc. World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education 2009 (E-Learn 2009). 26-30 October 2009, Vancouver, B.C., Canada (eds. Bastiaens, T. et al.). Association for the Advancement of Computing in Education (AACE), Chesapeake, Virginia, USA, 1741-1750. ISBN 1-880094-76-2. http://www.editlib.org/p/32712

Guided Generation of Pedagogical Concept Maps from the Wikipedia

Lauri Lahti
Helsinki University of Technology, Department of Computer Science and Engineering
P.O. Box 5400, FI-02015 TKK, Finland
Email: lauri lahti at hut fi

Abstract: We propose a new method for guided generation of concept maps from open access online knowledge resources such as Wikies. Based on this method we have implemented a prototype extracting semantic relations from sentences surrounding hyperlinks in the Wikipedia's articles and letting a learner to create customized learning objects in real-time based on collaborative recommendations considering her earlier knowledge. Open source modules enable pedagogically motivated exploration in Wiki spaces, corresponding to an intelligent tutoring system. The method extracted compact noun–verb–noun phrases, suggested for labeling arcs between nodes that were labeled with article titles. On average, 80 percent of these phrases were useful while their length was only 20 percent of the length of the original sentences. Experiments indicate that even simple analysis algorithms can well support user-initiated information retrieval and building intuitive learning objects that follow the learner's needs.

1 Introduction

There is a need for computational methods that enable generating automatically textual abstracts and graphical illustrations about various kinds of information. They are needed for example in education, decision-making, research and media. People are receiving an increasing flow of information daily and it has become a challenge to filter essential knowledge from the noise. Also modern working life insists attention to issues of increasing level of complexity. This poses requirements to find new ways to educate people and to equip them with appropriate tools to tackle with complex knowledge and to interpret and exploit it to its full extent. In computer-assisted education, a strong trend is to develop *learning objects* that are modular resources designed to explain learning objectives and *intelligent tutoring systems* that provide automated guidance like an experienced human tutor. For a pedagogically motivated and tailored learning experience, visualizations in many forms can support knowledge management (Eppler & Burkard 2006).

Various compact notation techniques, such as diagrams and flowcharts, are used to compress information to more manageable units and to highlight essential relations. However, punctuality becomes easily sacrificed and it is challenging to find a good balance with compactness and detailedness in visualization. We think that new domain-independent adaptive methods are needed to manage knowledge with a compact notation that has an optimal expressiveness. Interpreting compact notations is often easiest for people having a shared history although creative work benefits from varied backgrounds. We think that concept maps are an illustrative and adaptive notation technique that should be increasingly exploited to support collaborative creative work. *Concept maps* are graphs of labelled nodes that are linked with directed labelled arcs depicting the relations between the nodes.

Motivated by previous research we propose a new method for generating adaptive concept maps from open access online knowledge resources, such as Wikies. *Wikies* are web sites freely built and edited by a community of volunteers. Following the principles of our method we have designed and implemented a prototype application extracting semantic relations from the articles of *the Wikipedia* free online encyclopedia (http://en.wikipedia.org/). Corresponding to an intelligent tutoring system, this enables creating customized learning objects in real-time based on collaborative recommendations. Initial experiments have indicated good applicability of the method for collaborative creative work.

2 Previous research

2.1 Relatedness and ontologies

Despite the fast progress in neuroscience, some sufficiently competent models are currently needed to support efficient thinking. Thinking relies on abstract linguistic skills acquired and exploited in social knowledge sharing. Knowledge can be analyzed on various semantic levels and often the lowest level consists of concepts. According to a classical but criticized theory, concepts are structured mental representations that encode necessary and sufficient conditions for their application (Laurence & Margolis 1999). In computational natural language processing, the ambiguous mappings of words to concepts are often analyzed as correlation patterns in large text samples. Online knowledge resources have received increasing attention since they can be easily accessed and updated by anyone. In digital format related pieces of knowledge can be versatilely connected with hyperlinks. Semantic features of networks have been modelled from various perspectives including learning, graph-based representation and information flows (Gladun et al. 2007) (Baget et al. 2008) (Erétéo et al. 2009). Based on statistical analysis and probabilistic methods, models have relied on lexicographical resources like *WordNet* (Fellbaum 1998), manual statements like in *CYC project* (Lenat 1995) and the contents of the Wikipedia (Krötzsch et al. 2007). Anyway, it has remained challenging to extract automatically semantic knowledge from natural language documents.

Computational language models have used for example n-grams and hidden Markov models, as well as various tagging and parsing techniques. A common assumption has been that co-occurrence of certain words in small observation window and in specific order indicates their semantic relatedness and similarity. However, indexing word distributions from large corpuses typically results in sparse high-dimensional vector spaces that are often inefficient in making searches and comparing distances, despite of advancement in dimensionality reduction techniques. Categorization of documents often relies on weighting and ranking matching documents. Two basic trends are statistical and intelligent indexing. The former approach has suffered from unrealistic assumption of independence of the index terms. This has encouraged the latter approach which consists of conceptual and semantic indexing (Wang & Brookes 2004). *Text classification* has strongly relied on so called “bag of words” approach combined with for example k-nearest neighbour algorithms, support vector machines and neural networks. Thus, usually only words explicitly mentioned in the text fragments have been considered, assuming the vocabulary to be consistent everywhere. Knowledge resources used for creating classification models have often had a limited coverage and challenges to be updated. Also the agility to both generalize and differentiate has been limited. *Tf-idf weight* (term frequency – inverse document frequency) is a general statistical measure for evaluating how important a word is to an article in a collection of articles (Salton et al. 1988). It reaches high values if the word appears frequently in the article but rarely in the whole collection.

Network models enable many linking schemes to express parallel semantic relations between textual items on various levels of abstraction and to tolerate possibly overlapping and fuzzy categorization. In article networks, *PageRank* is a popular measure used to denote importance of an article based on the amount of arriving links and their corresponding value. An old interpretation is that the PageRank value of an article can express the chance that a random surfer will arrive to this article through a link (Page et al. 1999). Both tf-idf and PageRank measures have a limitation that to work well they initially need to perform a computationally heavy indexing through the collection of articles. One computational approach often referenced as “semantic web” relies on building a common model of knowledge, so called *ontology*, by defining simple relation statements that link concepts. There are challenges to ensure coherent categorization when combining statements from varied human contributors and to deal with ownership and neutral management policies of collaboratively built knowledge resources, such as *Open Directory Project* (Hammond et al. 2005). Maintaining a constant update rate can be difficult for many initiatives. Besides defining relation statements manually, ontologies can be extracted from web content labelled with community generated *tags* (Nauman et al. 2008). This metadata actively produced by bloggers and social bookmarking creates collections of *folksonomies*. However, loose coordination and non-explicit criterion induce ambiguity reflecting varied individual preference and experience. Abuse for search engine optimization and anonymity of collaborators can also reduce reliability of tagging.

2.2 Mining the Wikipedia

One promising domain to extract community generated tags for ontology construction is offered by the Wikipedia online encyclopedia. Constantly growing, it holds about 3 million articles in English (May 2009) providing an

actively updated cross-linked network of articles and statements. For example article titles, article categories and hyperlink texts can be exploited as they were "tags" of a Wikipedia article. They indicate keywords or keyphrases describing a natural language concept represented by the corresponding article. It is computationally favourable that many of these tag-like features in Wikipedia articles obey hierarchically evolving abstraction and facilitate identification of the most essential semantic relations. For example, only a fraction of words in an article are hyperlinks and the hyperlinks in the beginning of an article provide often definitive relations whereas later hyperlinks provide more illustrative and detailed relations. In addition, the hyperlink distribution in both basic and advanced articles usually inherently supports rising the abstraction level in reasonable steps when accessing hyperlinks. The presence of this layered abstraction in the hyperlink network of the Wikipedia is a critical feature that favourably supports building a true ontology. According to (Strude & Ponzetto 2006), collaboratively created folksonomy extracted from the Wikipedia can be used in artificial intelligence and natural language processing applications with the same effect as hand-crafted taxonomies or ontologies. They suggest computing semantic relatedness of concepts by retrieving corresponding Wikipedia articles and measuring their textual contents and paths in the category taxonomy.

(Gregorowicz & Kramer 2006) proposed a method for generating a robust term–concept network from Wikipedia to support tagging and information retrieval, assisted with a query language such as SPARQL. To decrease ambiguity they introduced three term categories: "actual concepts" correspond actual article pages, "alternative terms" correspond redirects and disambiguation pages, and "related concepts" correspond articles having bidirectional links with the current article page. They suggest developing rich ontological relationships between concepts by applying natural language understanding to the article contents. Instead of the full article text analysis, (Milne & Witten 2008) emphasized the use of hyperlinks and they disambiguate term–article mappings by exploiting three features: conditional probability, collocation and link distribution similarity. (Gabrilovich & Markovitch 2009) suggested representing natural language semantics in a high-dimensional space of concepts based on calculating tf-idf weights for corresponding Wikipedia articles. The semantic interpretation of a text fragment is combined from the vectors representing the individual words. Their model seemed to address well synonymy, polysemy and text excerpts of arbitrary length. Applying the model to the Wikipedia outperformed applying it to Open Directory Project, and comparing two temporal versions of the Wikipedia, the newer content gave a small but consistent improvement. Information obtained through inter-article links did not essentially improve computing pure semantic relatedness of texts. However, in text categorization which requires considering only few highest-scoring concepts in text fragments, the use of inter-article links improved accuracy. They suggest partitioning the article at several levels of linguistic abstraction (ranging from words to entire article) and generating features at each level. Inter-article links could be taken into account by boosting the weights of the concepts that are linked from a number of top-scoring concepts and giving preference for highly linked concepts.

An overview of knowledge mining from the Wikipedia is provided by (Medelyan et al. 2009). Knowledge mining from the Wikipedia has already been widely applied for various tasks, for example supporting validation of relevant information and combining various knowledge resources (Blohm et al. 2008) (Hoffmann et al. 2009). (Nakayama et al. 2008) showed that link structure mining improves both the accuracy and the scalability of semantic relations extraction from the Wikipedia. They propose three processes optimized for Wikipedia mining: fast pre-processing, part-of-speech tag tree analysis and mainstay (statement) extraction. Their method collects three different relation patterns from sentences of an article**:** normal ("is-a"), subordinate ("is-a-part-of") and passive ("was-born-in"), and assumes a subject to be followed by a verb and an object. To filter important sentences in the article they used a measure analogous to tf-idf weight. This *pf-ibf measure* (path frequency – inversed backward link frequency) is high for articles which share forward/backward hyperlinks with a domain-specific article but do not share with other articles. Based on previous principles, an online association thesaurus with a search engine and visualization has been implemented (Nakayama 2008). To identify if the subject of a sentence is same as the topic of the article they suggest three rules: the terms of the subject are all contained in the title, the subject equals with the most frequent pronoun of the article, or the subject matches with texts used in backward links of the article.

2.3 Customized learning objects

Building both learning objects and intelligent tutoring systems has typically been laborious and become cost-effective only in a highly specified domain. We wanted to find new ontology-related solutions. Unfortunately, among many defined *learning content models*, only part of them supports standardized ontology-based content and metadata (Verbert & Duval 2004) (Zouaq et al. 2007a). Since manual generation of ontologies is slow and prone to

errors, we considered automated or semi-automated methods as a necessity. Pedagogically, in information retrieval we are inspired by *question answering* that aims to return specific answers to questions. Exploiting this with the Wikipedia for fluent learning experience include already an interface for command line queries (Kaisser 2008), biography quizzes (Higashinaka et al. 2007), and a tool assisting Wikipedia authors (Jijkoun & de Rijke 2006). However, indication of promising learning paths, unconstrained exploration and intuitive visualizations are typically missing in current solutions. (Kumar 2006) argues that in intelligent tutoring systems managing domain models and learner models can get support from so called "domain concept maps". There does not currently exist many solutions supporting non-predefined verbal relations between concepts in the ontology and exploiting concept maps.

(Zouaq et al. 2007b) proposed a layered model that with natural language processing extracts concept maps from documents and organizes the generated knowledge into Web Ontology Language (OWL) document ontologies. Their method identifies document's keywords with surrounding sentences on the grounds of tf-idf weights and distance from the beginning. By extracting concept-verb-concept triples and other relations with a parser the method generates a semantic network which can be further modified by a human expert with a visual map editor. The method aims to find all verbal relations between concepts so that they can be later used in training process enabling varied pedagogical strategies. Two presentation modes are introduced: "concept map oriented mode" enables to access documents by selecting a concept and "learning object oriented mode" lets to build automatically learning objects that resemble the learner's needs and competence. Zouaq et al. suggest that the learner should be allowed to construct also her own concept maps dynamically from the list of available concepts and links.

We now suggest extending the use of ontologies and concept maps into semantic modelling with the supply of the Wikipedia. We propose that the collaboratively maintained knowledge structure of the Wikipedia can serve as a both adaptive and expressive frame for implementing customized learning objects. We propose extracting semantic relations from hyperlinks of an article and parsing compact explanations about them. The leaner, knowing herself best her needs, builds and explores in real-time this personalized content in the form of concept maps. A method listing few promising paths in this knowledge structure corresponds to a traditional intelligent tutoring system. Describing these paths visually will directly represent semantic knowledge in a compact, easily digestible form that is linked firmly with the most recently exposed knowledge. Our proposal should improve learning experience while decreasing complexity of technical implementation. We do not know other work similar to our proposal in respect to learner-driven generation of labelled concept maps extracted from Wikipedia hyperlinks. For example, Wikipedia Roll merely focuses on browsing hyperlinks grouped in article's subchapters (Muthesius et al. 2008). Outside the Wikipedia domain, resembling concept mapping efforts include (Dey et al. 2007) and (Nasharuddin et al. 2008).

Natural and social networks, including the Wikipedia, form hierarchical cluster structures even without human coordination. These structures emerge following so called power law in for example article sizes, the number of connecting links, editing times and collaboration distribution (Buriol et al. 2006). These structures support the network in minimizing average paths between nodes and maximizing ability to recover if a random node fails. We suggest that management of ideas and concepts in human mind and collaborative learning may rely on an analogous cluster structure and thus favourable learning paths could rest on experimenting with the hyperlink structure of the Wikipedia. Our proposed method tries to facilitate exploiting these cluster structure for educational purposes. In contrast with a traditional intelligent tutoring system, our approach does not require computationally extensive evaluation in the learning content space. Instead, a simple algorithm can ensure that the learner stays on collectively recommended learning paths and exploits intuitive linking between entities. Since perspective is at low conceptual level through narrative, the necessary requirement of continuity can be assessed well in a constructive manner. Since the knowledge in a Wiki platform is already initially organized following human intuition, there is no need for an intelligent tutoring system performing heavy mining to reformulate information and to interpret it to a human user. Even the choice between alternative learning paths can be given directly to the learner since the initial organization of knowledge and previous steps should be intuitive enough to support learner to make the best decisions for herself.

3 Concept maps from hyperlinks and their context

3.1 Method and prototype

The proposed method is based on extracting semantic relations from Wikipedia articles on the request of a learner and gradually building a concept map online representing learning paths following the learner's initiative and

interests. Evolving concept map provides functionalities of a customized learning object and an intelligent tutoring system that can be flexibly modified and reused. The prototype is available from the author on request. User interface of the prototype is shown in (Fig. 1). In the beginning, the learner should be provided with a general idea about fertile domains for exploration. When the learner has decided the topic of interest she needs to figure out one related start concept and add it manually as the first node of the concept map. By pressing the button "Suggest hyperlinks", the method retrieves a Wikipedia article having a title that matches the concept given by the learner. From the retrieved Wikipedia article the method extracts every hyperlink and the sentence surrounding it. A hyperlink consists of the title of the target article (title) and highlighted text sequence in the current article referring to the target article (anchor). For each extracted hyperlink the method generates a compressed explanation phrase based on the surrounding sentence.

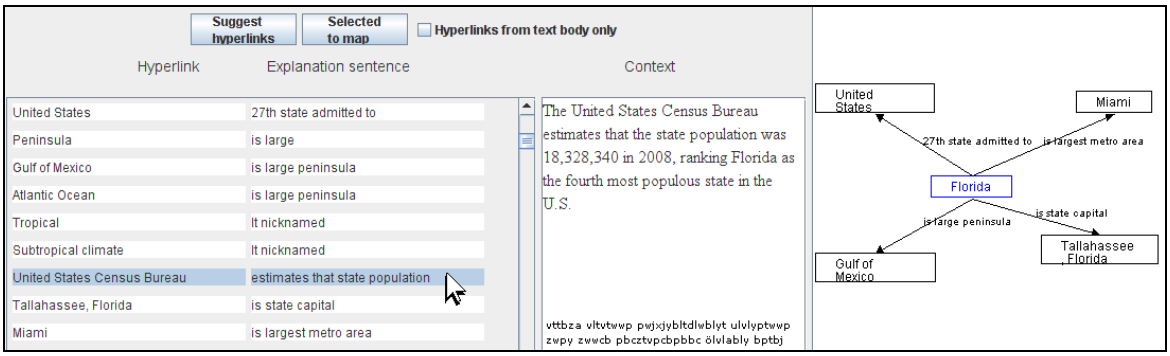

Fig. 1: User interface of prototype while exploring the hyperlinks of Wikipedia article about Florida (a detail).

Heavy compression is motivated by our aim to illustrate the semantic relation denoted by the hyperlink to the learner in a pedagogically most efficient and compact form consisting of only few words. Even if simplifying the original meaning, a reasonable compression seems to greatly support the learner to interpret and understand the relation in an educationally fluent way. This should address the limited space available to illustrate meanings in a concept map that is a medium trying to offer optimally condensed representations. The compression is done by identifying a verb and two adjacent nouns from the sentence and eliminating other less relevant words. This simple technique addresses effectively varying syntax in sentences. The hyperlinks are shown to the learner as a scrollable list in the original order of appearance thus promoting core definitions usually in the beginning of an article. Each row shows the title of hyperlink and its short explanation phrase. Following her intuition and evaluation, the learner can select one or more hyperlinks from the list. For the most recently selected hyperlink, the full original sentence is shown in a separate textbox to verify that the explanation phrase holds. By pressing the button "Selected to map" she can add them as new child nodes of the currently active node, connected with directed arc. The node label is derived from the title of hyperlink and the arc label from the explanation phrase respectively. After adding new linked nodes, they can be used as initial concepts for further exploration. Step by step the learner establishes and proceeds in the most promising learning path for her needs. (Fig. 1) illustrates operation when the topic of interest is "Florida". For example following hyperlinks and explanation phrases are provided: United States ("27th state admitted to"), Gulf of Mexico ("is large peninsula"), Tropical ("It nicknamed") and United States Census Bureau ("estimates that state population"). With common reasoning, the first two explanation phrases are correct, the third is incorrect and the fourth is apparently lacking actual relevance.

Besides exploiting the list of hyperlinks, the user interface allows the learner to freely add, rename and remove nodes or arcs in the map and drag them to the most illustrative positions. Due to ambiguity in semantic relation extraction, the learner is encouraged to verify the suggested explanation phrases in respect to the full sentence and her previous knowledge and to rephrase them if needed before adding them as new arc labels. Gradually expanding the concept map with new arc-connected nodes creates a graphical representation of learning paths following the learner's thinking and interests. Due to variety of authors highlighting different aspects in the articles our proposed method enables the user to gain new perspectives with surprising associations and therefore to be more creative and innovative. Also the proposed method helps the user to keep track of learning paths she has taken. The method lets the user to see with an intuitive and compact graphical notation both the earlier choices and the current alternatives to make consent decisions in reasoning. We think that maintaining this constant holistic view supports reaching

pedagogically motivated and innovative learning paths. By saving the generated concept maps they can be easily applied as a specification for learning objects since directed arcs connecting nodes correspond naturally relations of causality or hierarchy in curriculum. The maps can be also used later as a reference of earlier reasoning and enabling evaluation of learning progress since then. The link structure and labels of the map can be iteratively edited and updated to address current learning needs. Various alternative maps can be created about a topic by selecting different learning paths and these maps can be compared to induce recommendation for future.

3.2 Modular implementation

We have implemented a functional prototype that consists of independent modules and uses parallel processing. We think that to accelerate development and use of advanced educational applications it is necessary to foster creative reuse of available open source components. Instead of closed architectures, developers should actively share models and algorithms. However, to ensure interoperability defining universal interface standards should have a high priority. In the platform handling concept maps we continue work of (Lahti & Tarhio 2008) but instead of gaming with the fixed hierarchy of a table of contents we now enable the learner to build and explore customized learning objects in real-time. Four main tasks of our method are retrieving appropriate Wikipedia articles, extracting hyperlinks with surrounding sentences, tokenizing the sentences with part-of-speech tags and generating explanation phrase for each hyperlink based on the surrounding sentence. We use *Apache Commons HttpClient* module to submit queries and to retrieve articles from the Wikipedia (Apache Commons 2009). We use *CRFTagger* module developed by Xuan-Hieu Phan as a module of part-of-speech tagging for English. According to documentation, it is based on first-order Markov conditional random fields model trained on Wall Street Journal portion of the Penn Treebank corpus and is said to achieve accuracy of 97 percent (Phan 2006). Compared with rule based or lexicon dependent approach, a typical advantage of a Markov model is adaptivity to varying lexical contexts although at the cost of some accuracy. We designed and implemented the algorithms that extract hyperlinks with surrounding sentences and generate explanation phrases, and the algorithm coordinating existing modules.

Extraction of hyperlinks with sentences is based on scanning through the html encoded Wikipedia article file character by character. The borders of a hyperlink are recognized by sequences <a href="/wiki/ and . The sentence borders are basically recognized by a dot followed by a space and a capital letter, or by tags declaring a new entity. When reaching the end of a sentence, it is stored together with all of its hyperlinks. With a retrieved article CRFTagger module first applies Penn Treebank part-of-speech conventions to distinguish special notations. Based on the trained model of CRFTagger each text item is tokenized with a tag representing its role in the sentence. Generating a compact explanation phrase from the tokenized sentence relies on identifying a noun–verb–noun sequence closest to the hyperlink anchor. Nouns are typically words or word groups that can serve as the subject or the object of a verb. The closest nouns on the left side and the right side of the verb define the borders of an expression that will be compressed further. The algorithm tries to eliminate unnecessary words from the expression, such as possible occurrences of article title and hyperlink anchor, auxiliary verbs "can/may/be/have", articles "a/an/the" and parts of nouns following an "of". In the minimum, an explanation phrase contains only the hyperlink anchor. To achieve a high compression of the explanation phrase with a simple algorithm we decided to accept loosing essential details occasionally. Anyway, we were impressed how well a human observer can tolerate fragmented information and spontaneously complement biased explanation phrases to still catch the original meaning. If the hyperlinks and their explanation sentences for a certain node in the concept map have been already generated during the current session, they are later accessed from local cache by just activating this node by mouse. To reduce consumption of bandwidth, occupation of Wikipedia servers and overall computation, we consider this solution sustainable even if loosing a possible sudden update in the Wikipedia article. Depending on the article size, producing hyperlinks and explanation phrases with our method had varied duration and took on average one minute with a modern PC but the rise of computing power will rapidly reduce this cost.

4 Experiments with the prototype

In early May 2009 we conducted experiments with our prototype by generating concept maps from the Wikipedia and evaluating their pedagogical quality with human reasoning. From a listing of 1000 most visited articles of the Wikipedia in 2008 (Falsicon 2009) we randomly retrieved 20 articles and automatically generated an explanation phrase for each hyperlink they provided. The articles covered diverse topics including four countries, three movies, three music groups and four person-related articles which motivates us to generalize the results. We decided to limit

our analysis on the introductory text paragraphs appearing before the table of contents since these sections obviously are the most frequently read parts of the article and try to represent the topic in a condensed form. We extracted a total of 543 hyperlinks ranging from 5 to 48 per article. The sentences surrounding these hyperlinks had an average length of about 25 words. The generated explanation phrases had an average length of 4,2 words meaning compression by over 80 percent from the length of the original sentences surrounding the hyperlinks. When evaluating the pedagogical quality of generated explanation phrases we manually classified them into three separate categories based on their intuitiveness (Tab. 1). A "useful phrase" describes the relation between the current article and the hyperlink with a truthful expression that provides some new semantic value. A "misleading phrase" carries a similar kind of semantic expression that is convincing but deceptively false. A "fuzzy phrase" offers an expression lacking true semantic value but anyway is so confusing that it is rather easy to identify and ignore.

In initial evaluation, it became apparent that even pretty coarse explanation phrases could reliably convey true and valuable meaning. A typical ambiguous case was when concepts had been extracted from a list of alternatives mentioned in the text. For example, an article about "Lion" provided a hyperlink "Painting" with explanation phrase "It depicted in literature". We considered here that the learner should intuitively manage to replace in explanation phrase literature with painting, recognizing that both are forms of art. We admit that requiring this much semantic tolerance is a drawback for usability but suggest that it is still manageable ambiguity for ordinary learners. Another ambiguous case was when the explanation phrases did not specify the direction of a relation. For example, an article about Italy provided hyperlink "European Union" with explanation phrase "member what is". Here again, we assume that the learner can intuitively manage to interpret direction correctly. For all articles together, 81 percent of explanation phrases appeared to be useful, 11 percent misleading and 8 percent fuzzy. Only exception to the general success rate of at least 69 percent, is article about "Filippines" with 33 percent success only. A closer look revealed this being apparently due to having a lot of sentences referring to various cultural and geographical concepts that the method could not succeed in mapping correctly. We consider current success rate surprisingly good and convincing, especially in respect to high compression of explanation sentences. An extensive result list of the experiment is available at the author's web site (http://www.cs.hut.fi/u/llahti/publ/lahti_2009b_data.pdf).

	Lion	Italy	Radiohead	Jesse McCartney	Hancock (film)	Pink Floyd	Bow Wow	Philippines	Judaism	John Cena	Kenya	Linux	Star Wars: The Clone Wars (film)	Iraq War	Terminator Salvation	ABBA	Florida	Two-Face	Religion	Peru	∑
Explanations	38	48	29	5	18	32	12	18	46	29	15	23	21	39	18	27	15	36	42	32	543
- useful	33	41	24	4	15	22	11	6	39	25	14	16	15	28	15	26	12	29	38	26	439
- misleading	4	5	3	0	2	5	0	7	5	4	1	4	3	4	1	1	2	3	3	5	62
- fuzzy	1	2	2	1	1	5	1	5	2	0	0	3	3	7	2	0	1	4	1	1	42
Success	87 %	85 %	83 %	80 %	83 %	69 %	92 %	33 %	85 %	86 %	93 %	70 %	71 %	72 %	83 %	96 %	80 %	81 %	90 %	81 %	81 %

Tab. 1: Distribution of useful, misleading and fuzzy explanation phrases generated for hyperlinks of twenty Wikipedia articles separately. The success percentage indicates the proportion of useful phrases to all phrases.

5 Discussion

We think that the proposed method can facilitate pedagogically motivated knowledge management in many ways. The method relies on a constantly growing and collaboratively fine-tuning large online knowledge resource, the Wikipedia. The method supports a learner to explore independently the densely cross-linked pieces of up-to-date knowledge following spontaneously her own educational needs. By extracting relations from hyperlinks of the Wikipedia the method illustrates intuitively the most promising learning paths based on recommendations given by a diverse community. The learner can build and experiment with compact visualizations that represent her understanding and taken perspectives. Resulting concept maps indicate clearly the relations of facts supporting constructive learning paradigms and creating sustainable customized learning objects. The learning process is inherently self-regulating since previous learning paths and the most probable future directions are efficiently presented and comparable all the time. Evaluating various perspectives with a critical attitude is well supported. Proceeding in the learning content space can be performed with manageable steps in abstraction level and minimizing excessive cognitive load. All concept maps built by an individual learner can be agglomerated to greater

entities and used as customized learning objects. The method is flexible since it can be applied equally well to exploring details of a specific domain or to ideation of distant associations. The method addresses typical requirements for creative problem solving providing surprising viewpoints yet enabling sustainable continuity to old knowledge. We think that the method can be implemented in a variety of tools to assist individual and collaborative learning process in a systematic manner. As an example, we expect that the proposed method could be integrated into the platform suggested in our previous work (Lahti, 2009).

In the proposed method learning efforts become well documented and the produced visualizations can be easily reused, updated and shared. By tracking the building process of concept maps, teachers can practically evaluate learning progress in respect to learner's individual resources. The method also enables teachers to update their own knowledge and plan curriculum. By analysing the temporal construction phases of a concept map can assist identifying and responding to various learning styles. With small modifications the method could be transformed to generate automatically concept maps for school lessons with a great variation and always up-to-date. These concept maps could be tailored to address varying topics and learning styles of each attending learner. Extending the method to parallel language versions of the Wikipedia or other Wikies could enable new ways to understand cultural and language related differences in conception and ontologies. In addition, learning foreign languages could be supported with comparison of conceptual relations simultaneously in two language versions. Furthermore, in special education and assistive tools various everyday processes could be illustrated. The method can supply information retrieval and question answering with close personalized touch. A great diversity of easily digestible pieces of knowledge can be provided to the learner with the method. Even if the learner is challenged in her cognitive skills, the method still guarantees her rights to make the ultimate decisions about the learning path to proceed. Besides text, the concept maps could be easily transformed to exploit multimedia content. In addition, various metrics could be applied to assist the learner to identify the most mature and trusted content in the online knowledge resource. Thus the method could promote using the most extensive and reliable learning paths. Even if the method occasionally provides inaccurate knowledge it can be exploited as a learning resource that urges the learner to critically evaluate the content and make rephrasing that is well mapped to her previous conception. Incomplete explanation phrases offered to support building concept maps can be considered as a valuable way to activate the learner to excel oneself in personal knowledge acquisition and formulation. Completing the phrases can be used as a personalized exercise to evaluate learning progress so far. The learner becomes actively encouraged to rephrase the relations suggested by the method so that they fully correspond to her own intuition.

In contrast with many other research proposals in this field, we have implemented a fully functional prototype and with experiments verified the success of our proposed method. We think that too often educational practices rely on unverified beliefs. We want to actively promote bringing theoretical research results into everyday school environment to increase productivity and quality of life. Due to modular structure, the functionality of our method can be flexibly extended and modified later to exploit new better modules following the latest pedagogical insight. We also think that the patterns of learning emerging in school life should be exploited much more to develop new theoretical models. The proposed method and the related prototype indicate new possibilities to facilitate tracking learning events at schools to find better models to support learning. Long-term studies with large populations are needed to better understand the long-lasting and slowly evolving learning processes in individual minds. It is possible that earlier research has too optimistically aimed at single models that could favourably support different learners. We think that the proposed method can give directions for new learning practices that evolve and mature together with each individual learner. For example, curriculum and learning objects may often be too fixed and aimed at an average learner only. To liberate education from too homogenous one-for-all standards we need to cope with challenges of identifying the great variety in the learning progresses of individuals. To really address all learning difficulties it is a necessity to take into account different personalities, temperaments and interaction styles acquired during the early childhood. Increasing personal knowledge and educational level should be seen as an important goal for everyone, affecting only positively to well-being.

We think that the proposed method offers practices to be considered as mediators to enhance understanding individual learning styles and how they are related to educational needs. To capture the essence of the holistic learning process performed by an individual mind requires new analytical approach that should increasingly exploit latest technology, such as information networks, mobile communication and virtual teams. In school, the educational practices should aim to provide life-long learning skills not only based on today's requirements but also trying to predict tomorrow's requirements. To stay in the first wave, it is important to model how new knowledge can be submerged with prior knowledge and how rich adaptive representations can support this process. We think that

learner-driven unconstrained experimenting with various conceptual structures can be a key factor in the development of new advanced support tools. It seems to us that extensive indexing of knowledge from online resources before a learning process has even started cannot fully satisfy the individual needs of a learner. We think that the learner should get thoughtful guidance but eventually to be free to make creative initiatives following her intuition. We think that exploration patterns should be well documented so that they could be directly exploited in building collective knowledge structures, beneficial for other learner's later as well. Along the years, learning process of an individual should produce conceptual structures that illustrate her core understanding, like an autobiography in a form of a visualized relational database.

We aim to develop further the pedagogical advantages of our proposed method. The method can be extended to retrieve automatically concept maps about a wide range of topics to provide ready-made learning objects. These maps could be used as an augmented user interface for browsing the Wikipedia. Even in offline mode the maps could serve as a compact search tool representing conceptual relations since many fundamental facts are fixed and do not change daily. With a shared online platform individuals could use the methodology to build mutually agreed concept maps. This should support constructive dialogue to find resolution ensuring that all opinions become heard. We think that it is important that our method supports drawing concept maps even without retrieving knowledge from the Wikipedia. In the case that the Wikipedia is temporarily inaccessible or it provides irrelevant or false information it is important that the user can freely decide the structure and labelling of the concept map. Based on promising results in initial experiments, we are planning to carry out wider user testing in collaborative environment. We think that the proposed method shows how important it is to support free exploration in conceptual spaces and recognize many equally valid alternative conceptions. We think that learning through trial and error can well support iteratively refining processes of human thinking. Future research should give attention to modelling how the construction of pedagogically favourable concept maps really relies on the features of unrestricted exploration. Thus there is a need to explain how the learner actually can benefit from experimenting with the keywords of a learning topic in a concept map following her intuition. Recommendable practices of knowledge management should be indentified and used for developing new adaptive tools that support learning, innovation and creative problem solving. Domain-independent methods to explore knowledge and represent it illustratively should have a high priority in the research agenda.

References

Apache Commons (2009). HttpClient project. http://projects.apache.org/projects/commons_httpclient.html

Blohm, S., Krötzsch, M., Cimiano, P. (2008). The Fast and the Numerous – Combining Machine and Community Intelligence. Proc. AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, Technical Report WS-08-15.

In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, volume Technical Report WS-08-15. AAAI Press, July 2008.

Baget, J., Corby O., Dieng-Kuntz, R., Faron-Zucker, C., Gandon, F., Giboin A., Gutierrez, A., Leclère, M., Mugnier M., Thomopoulos, R. (2008). Griwes: Generic model and preliminary specifications for a graph-based knowledge representation toolkit. Proc. International Conference on Computational Science (ICCS'2008).

Buriol, L., Castillo, C., Donato, D., Leonardi, S., & Millozzi, S. (2006). Temporal analysis of the Wikigraph. Proc. IEEE/WIC/ACM International Conference on Web Intelligence, 45–51.

Dey,L., Abulaish, M., Goyel, R., & Jahiruddin (2007). Semantic integration of information through relation mining – application to bio-medical text processing. LNCS 4815, 365-372.

Eppler, M., & Burkard, R. (2006). Knowledge visualization – towards a new discipline and its fields of application. In David G. Schwartz (ed.), Encyclopedia of Knowledge Management. Idea Group Inc.

Erétéo, Guillaume and Gandon, Fabien and Buffa, Michel and Corby, Olivier (2009) Semantic Social Network Analysis. In: Proceedings of the WebSci'09.

Falsicon (2009). Page hits per day for en.wikipedia.org in year 2008. Based on 210 analysed days, requests counted by Squid servers. http://wikistics.falsikon.de/2008/wikipedia/en/ (accessed May 2009)

Fellbaum, C. (ed.) (1998). WordNet – an electronic lexical database. MIT Press.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34, 443–498.

Gladun, A., Rogushinab, J., García-Sanchezc, F., Martínez-Béjarc, R., & Fernández-Breisd, J. (2007). An application of intelligent techniques and semantic web technologies in e-learning environments. Journal of Expert Systems with Applications, 36, 1922-1931

Gregorowicz, A., & Kramer, M. (2006). Mining a large-scale term-concept network from Wikipedia. Technical report. MITRE Corporation, Bedford, MA, USA.
Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005). Social bookmarking tools (I) - a general review. D-Lib Magazine 11(4).
Higashinaka, R., Dohsaka, K., & Isozaki, H. (2007). Learning to rank definitions to generate quizzes for interactive information presentation. In Companion volume to Proc. 45th Annual Meeting of the Association for Computational Linguistics, 117–120.
Hoffmann, R., Amershi, S., Patel, K., Wu, F., Fogarty, J., Weld, D. (2009). Amplifying community content creation using mixed-initiative information extraction. Proc. Conference on Human Factors in Computing Systems 2009.
Jijkoun, V., & de Rijke, M. (2006). Overview of the WiQA task at CLEF 2006. Proc. 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006), LNCS 4730, 265–274.
Kaisser, M. (2008). The QuALiM question answering demo: supplementing answers with paragraphs drawn from Wikipedia. Proc. Annual Meeting of the Association for Computational Linguistics combined with the Human Language Technology Conference (ACL-08 HLT), Demo Session, 32–35.
Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R. (2007). Semantic Wikipedia. Journal of Web Semantics. 5, 251--261.
Kumar, A. (2006). Using Enhanced Concept Map for Student Modeling in Programming Tutors. Proc. Florida Artificial Intelligence Research Society Conference (FLAIRS 2006).
Lahti, L. (2009). Assistive tool for collaborative learning of conceptual structures. Proc. Human Computer Interaction International 2009, Vol. 3, LNCS 5616, 53–62.
Lahti, L., & Tarhio, J. (2008). Semi-automated map generation for concept gaming. Proc. IADIS International Conference Gaming 2008 (part of MCCSIS 2008), 36-43.
Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In Concepts: Core Readings, MIT Press, 3–81.
Lenat, D. (1995). Cyc: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11).
Milne, D., & Witten, I. (2008). Learning to link with Wikipedia. Proc. ACM Conference on Information and Knowledge Management (CIKM'2008).
Medelyan, O., Milne, D., Legg, C., & Witten, I. (2009). Mining meaning from Wikipedia. International Journal of Human-Computer Studies. 67(9), 716–754.
Muthesius, T., Legois, D., Ramus, C., & Bourdu, S. (2008). Wikipedia-Roll browsing application. http://api-exploration.net./mashups/wikipedia-roll/index_en.php
Nakayama, K. (2008). Extracting structured knowledge for semantic web by mining Wikipedia. Proc. 7th International Semantic Web Conference (ISWC2008), Posters & Demos.
Nakayama, K., Hara, T., & Nishio, S. (2008). Wikipedia link structure and text mining for semantic relation extraction – towards a huge scale global Web ontology. Proc. SemSearch 2008, CEUR Workshop, 59–73.
Nasharuddin, N., Hamid, J., Ibrahim, H., Selamat, M., Abdullah, R., & Isa, W. (2008). Visualizer for concept relations in an automatic meaning extraction system. VINE: The journal of information and knowledge management systems, 38 (2), 232–240.
Nauman, M., Khan, S., Amin, M., & Hussain, F. (2008). Resolving lexical ambiguities in folksonomy based search systems through common sense and personalization. Proc. SemSearch 2008, CEUR Workshop. 2–13.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web. Technical Report SIDL-WP-1999-0120. Stanford University.
Phan, X. (2006). CRFTagger: CRF English POS Tagger, developed by Xuan-Hieu Phan, Graduate School of Information Sciences, Tohoku University. http://crftagger.sourceforge.net/.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. International Journal of Information Processing and Management 24 (5), 513–523.
Strube, M., & Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. Proc. National Conference on Artificial Intelligence (AAAI-06), 1419–1424.
Verbert, K., & Duval, E. (2004) Towards a global architecture for learning objects: a comparative analysis of learning object content models. Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications, 202–209.
Wang, B., & Brookes, G. (2004). A semantic approach for Web indexing. Proc. Sixth Asia Pacific Web Conference, LNCS 3007, 59–68.
Zouaq, A., Nkambou, R., & Frasson, C. (2007a) An integrated approach for automatic aggregation of learning knowledge objects. Interdisciplinary Journal of Knowledge and Learning Objects, Vol. 3.
Zouaq, A., Nkambou, R., & Frasson, C. (2007b). Document semantic annotation for intelligent tutoring systems: a concept mapping approach. Proc. Florida Artificial Intelligence Research Society Conference (FLAIRS 2007).