# PERSONALIZED LEARNING PATHS BASED ON WIKIPEDIA ARTICLE STATISTICS

Lauri Lahti
*Helsinki University of Technology (TKK), Department of Computer Science and Engineering, Finland*

Keywords: Semantic navigation, Intelligent tutoring system, Concept map, Wikipedia, Ranking.

Abstract: We propose a new semi-automated method for generating personalized learning paths from the Wikipedia online encyclopedia by following inter-article hyperlink chains based on various rankings that are retrieved from the statistics of the articles. Alternative perspectives for learning topics are achieved when the next hyperlink to access is selected based on hierarchy of hyperlinks, repetition of hyperlink terms, article size, viewing rate, editing rate, or user-defined weighted mixture of them all. We have implemented the method in a prototype enabling the learner to build independently concept maps following her needs and consideration. A list of related concepts is shown in a desired type of ranking to label new nodes (titles of target articles for current hyperlinks) accompanied with parsed explanation phrases from the sentences surrounding each hyperlink to label directed arcs connecting nodes. In experiments the alternative ranking schemes well supported various learning needs suggesting new pedagogical networking practices.

## 1 INTRODUCTION

To enhance the quality and efficiency of automated information processing there is a need for new computational methods. The methods used for representing and modifying information with the computers should be made more compatible with the methods that are naturally used by humans to adopt and associate meanings. Thus developing illustrative adaptive computational methods that can be used in knowledge management in natural language and with intuitive visualizations should have a high priority in research agenda.

We suggest that information available in structured online knowledge bases can serve as a valuable resource for computer-assisted learning in respect to the key concepts of curriculum and semantic relations linking them. We propose a new adaptive method to assist individual learning of core network of semantic relations in curriculum and later expanding this network further. We consider that still today the most valuable environment for learning is social collaboration in everyday life. We do not aim to challenge this traditional method but to complement it in a fruitful manner. We propose a new semi-automated method for generating personalized learning paths from the Wikipedia online encyclopedia by following inter-article hyperlink chains based on various rankings that are retrieved from the statistics of the articles. A *learning path* describes a structure of actions a learner has to perform in order to attain a competence or a competence profile (Janssen et al., 2008). In our proposal the learning paths are represented with concept maps. We are interested in methodology related to semantic navigation, intelligent tutoring systems and content-based filtering.

Our method aims to ensure that an optimal coverage of concepts becomes provided to the learner. Especially our method aims to enable active explorations of conceptual relations taking into account various diverse perspectives that are based on different individual backgrounds and preference. Besides in education, the method can be used in various professions to support creative problem solving to analyse various perspectives and to acquire distant associations for inspiration. The method can be further applied to model cultural dependencies in conceptual structures and how they can be flexibly exploited in cultural exchange for better mutual understanding.

## 2 BACKGROUND

Among many competing learning theories *constructivism* has remained widely supported. In



Lahti L. (2010).
PERSONALIZED LEARNING PATHS BASED ON WIKIPEDIA ARTICLE STATISTICS.
In *Proceedings of the 2nd International Conference on Computer Supported Education*, pages 110-120
DOI: 10.5220/0002800901100120

brief, it states that humans generate knowledge and meaning from their experiences. Holmes et al. (2001) suggested an expanded definition of social constructivism that could fully address the synergy between advances in information technology and virtual environments. One general challenge comes from the long-lasting debate if semantic structures of natural language are independent of syntactic structures or not (Peregrin, in press). *Transferable learning* that enables applying previously acquired training successfully for novel future events can be achieved through the learner being exposed to the learning material in a variety of contexts (Schmidt & Bjork, 1992). Designing learning activities can exploit the notion that people typically predict upcoming words in fluent discourse (Van Berkum et al., 2005). There is evidence that concept-oriented reading instruction increases reading comprehension and engagement (Guthrie et al., 2004).

Collaboratively maintained web sites, *wikies*, have been actively adopted as new educational environments with an assumption to support constructive learning process. However, typical use of wikies may enhance merely student engagement, but not performance on assessment (Neumann & Hood, 2009). A leading wiki site, *the Wikipedia online encyclopedia*, provides an extensive coverage of factual knowledge from various domains of life and is actively used as a resource by students and educators. Despite the concerns of accuracy, missing reference and vandalism, the content has been shown to be relatively reliable and up-to-date (Chesney, 2006). The content can be added and edited collaboratively by anyone but some parts are more protected to prevent vandalism and consistent rewriting. General usage patterns for various Wikipedia editions have been analyzed (Reinoso et al., 2009) showing a ratio of 620 reading operations per one saving operation for articles in the English edition.

It has been estimated that in English lexicon there are well over 54 000 word families and an educated adult native speaker knows around 20 000 of them (Nation & Waring, 1997). However, the most frequent 3000 to 5000 word families typically cover around 90 % of ordinary text and even more of spoken language of a language. Mastering just this fraction can already provide a strong basis for comprehension thus allowing efficient further learning from the context. On the other hand, the Wikipedia online encyclopedia currently contains over 3 million articles in English thus greatly exceeding the average vocabulary of an educated adult. This motivates us to propose exploiting the Wikipedia as a valuable resource for developing methods that assist people in adoption of conceptual relations and learning in general.

There have been attempts to develop methods to generate personalized representations from large knowledge resources to serve for example in education. After introducing our method and initial experiment, we will discuss about some related work in Section 5.

# 3 METHOD

## 3.1 Ranking Hyperlinks based on Article Statistics

Depending on the Wikipedia article, the amount and type of hyperlinks that it holds varies a lot. The more hyperlinks exist, the more alternative learning paths can be provided to the learner although making it also harder to choose one of them through comparison. A fundamental computational challenge is to identify the most promising hyperlinks and indicate them to the learner. For natural language processing applications, various confidence measures have been developed to estimate the probability of correctness of the outputs (Gandrabur et al., 2006).

We propose that the hyperlinks can be shown in a list that supports ranking based on various criteria taking into account different perspectives provided by each hyperlink's target article and depending on varied preferences among learners. Obviously, using many parallel measures for ranking hyperlinks can enhance possibility to systematically differentiate alternative rankings but unfortunately also increases computational complexity. It could be also possible to perform deep searches in the network and based on them conclude the most promising direction to traverse the next hyperlink. However, we wanted to minimize the cost of searches in the network and decided to evaluate only those articles that can be reached within a distance of one hyperlink step.

Since we aim at developing methods that can be used even with modest technological resources, we want to consider now simple measures only. As computational power constantly grows we expect taking increasingly complex measures into use in the future.

We propose that many statistical features about the hyperlink's target article can be retrieved as useful indicators about the perspectives that the target article represents in relation to the current article. Thus, one can use alternative ranking principles to sort target articles of hyperlinks in the

current article. This enables getting target articles of hyperlinks to be promoted in varying order of preference, depending on to which statistical features have been given priority in ranking. The highest-ranking hyperlinks with each alternative ranking can provide different perspectives for exploration in the article network for a learner.

In our work we decided to generate ranking of hyperlinks with such simple features that would be as much as possible motivated by the main functionality of the Wikipedia. We concluded by naming five key functions of the Wikipedia and corresponding measurable features for ranking. They are: adding new content (*article size*), editing content (*editing rate*), providing cross-linking (*hierarchy of hyperlinks*), explaining concepts and their relation (*repetition of hyperlink terms*) and using articles as a reference (*viewing rate*). Each of these five features enable relatively straightforward ranking of hyperlinks.

For example in November 2009, Wikipedia article about "Life" contained hyperlinks to target articles such as "Earth" (9152 edits, size 417 499 bytes) and "Metabolism" (1478 edits, size 456427 bytes). With common statistical reasoning, one could expect that the coverage of the article "Metabolism" (bigger article) might be broader but that the peer-review process of the article "Earth" (more edits) might be more extensive. Among these two hyperlinks, prioritizing the article size would promote relating concept of life to metabolism and prioritizing the number of edits would promote relating concept of life to earth. Naturally, cautiousness is needed in statistical evaluation especially when having relatively low frequencies.

From the sentence surrounding the hyperlink it is possible to parse and extract a compact explanation phrase that depicts the semantic relation between current article and the hyperlink's target article. In the previous example, one could produce two alternative robust relation statements from the text of article "Life" (November 2009): that forms of life "can be found in the biosphere on earth", and that during life "living organisms undergo metabolism".

For a learner to exploit the perspective that is currently emphasized in ranking, it is recommendable to traverse some of the high-ranking hyperlinks. The current article, the target article corresponding to the selected high-ranking hyperlink and the relation statement between them can be intuitively represented and further evaporated in a *concept map*. In the concept map, nodes labelled with article titles are connected with directed arcs labelled with relation statements. By expanding the concept map step by step with a preferred ranking of hyperlinks the learner can explore and build learning paths emphasizing desired perspectives.

## 3.2 Principles of Ranking

We think that the order of appearance of hyperlinks in the article is the simplest ranking of hyperlinks to exploit since it is inherently available in the article text. This ranking can be assumed to suggest that the hyperlinks in the beginning of the current article point to articles whose titles emphasize giving a definition of the current article. Reason for this is that a Wikipedia article often starts with a compact definition containing a few hyperlinks. Respectively, the hyperlinks in the end of current article likely point to articles whose titles emphasize giving broader details of the current article. We will refer to this type of ranking as the "Hierarchy of hyperlinks". Other rankings can be expected to rely on alternative prioritization for hyperlinks.

Statistical features of an article can be computed directly from the article or its revision history, or then retrieved from the open statistics database provided by the Wikipedia foundation. Several specialized web sites provide an easy interface for making queries with the statistics database.

In preliminary testing we evaluated a varied randomized sample of 100 Wikipedia articles. We tried to identify what kinds of target articles of hyperlinks become typically favoured when ranking is performed in respect to each of five features introduced in Section 3.1. We also tried to identify cases in which some articles become misleadingly favoured against these expectations just noted for each ranking. Table 1 shows our conclusion based on the sample and describing a hypothesis about favourable and misleading cases. We assume that on average same kinds of tendencies appear with any random Wikipedia article when ranking its hyperlinks in respect to each of five features.

We propose a new method that enables to generate personalized learning paths from the Wikipedia online encyclopedia based on various rankings that are retrieved from the statistics of the articles. The method should fruitfully support principles of constructivism and transferable learning.

Relying on the current learning task the learner needs to be first provided with a start concept for exploration. The method retrieves a Wikipedia article whose title matches with this concept. Then the learner is provided with a list of hyperlinks in this current article, sorted in a desired type of ranking based on the statistics of target articles. The list shows the title of target article for each

Table 1: Some favourable and misleading cases for ranking articles identified in respect to five features.

| | Hierarchy of hyperlinks | Repetition of hyperlink terms | Article size | Viewing rate | Editing rate |
|---|---|---|---|---|---|
| *Favour-able cases:* | - compact definitions in the beginning<br>- later illustrative and more detailed uses, alternatives | - everyday vocabulary<br>- general topic with many variations and sub-branches | - key terms of each field<br>- stabilized knowledge, biographies | - recent news topics, trends in popular culture<br>- technology, entertainment, celebrities | - controversial, non-stabilized or actively evolving<br>- science, politics |
| *Mis-leading cases:* | - any complex term that needs explanation<br>- unnecessarily broad or general terms | - use of synonyms or it/this hides the terms<br>- long terms less likely to be repeated | - single author's devotion without general interest<br>- article not condensed or yet split | - tourist information<br>- checks for equations, minor facts or spelling | - target of vandalism or consistent rewriting<br>- translated article suffering from low rate |

hyperlink, accompanied with a compact relation statement parsed from the sentence surrounding the hyperlink. The learner is expected to evaluate the provided hyperlinks and select intuitively, based on personal needs and consideration, the most promising hyperlink for extending exploration in the article network. By selecting especially a high-ranking hyperlink the learner can emphasize getting a perspective specific to present type of ranking. Each selected hyperlink progressively expands a concept map that is shown to the learner, defining learning paths. According to the selected hyperlink, a new node is added to concept map, labelled with the title of the target article. A directed arc is added correspondingly leading from current node to new node, labelled with the relation statement.

According to personal needs, the learner can choose which type of ranking is used for sorting the hyperlinks. The hyperlinks are sorted based on five different rankings that are generated from the statistics of the target articles. Alternative perspectives are available based on following five features, introduced in Section 3.1. As we have already motivated earlier, *"Hierarchy of hyperlinks"* denotes showing hyperlinks in the natural order of increasing distance from the beginning of the article. *"Repetition of hyperlink terms"* denotes showing hyperlinks in a descending order based on how many times the word (or group of words) forming the title of hyperlink's target article is mentioned in the current article, anywhere in its full textual content. This ordering is motivated by an assumption that the title of target article for each hyperlink defines a key term for current article. The more this key term is repeated in the text of current article, the more it seems to indicate that the corresponding target article is highly involved in formulating relations with the current article. *"Article size"* denotes showing hyperlinks in a descending order based on the total amount of characters in the target article text. A motivation for this ordering is that a bigger article size obviously indicates more detailed content than a smaller article size. The value of article size is approximated with the file size in bytes that is extracted from the header of the target article file.

*"Viewing rate"* denotes showing hyperlinks in a descending order based on frequency of viewing hyperlink's target article by the global community. This ordering is motivated by the assumption that an article with a high viewing rate has a higher general interest than an article with a low viewing rate. This value represents total number of views per previous full month. This information is retrieved from online service (Wikipedia article traffic statistics, 2009) that relies on data gathered from Wikipedia's squid-based cache server cluster. *"Editing rate"* denotes showing hyperlinks in a descending order based on frequency of editing hyperlink's target article by the global community. A motivation for this ordering is that higher editing rates seem to indicate more verified content than lower editing rates. The value of editing rate is approximated with the total number of edits for current article since its creation. This information is retrieved from online service (Wikipedia page history statistics, 2009) that builds an edit history overview page for the article with the given name.

Besides these five principal features, we still suggest a supplementing feature that is a user-defined weighted mixture of them all.

# 4 EXPERIMENT

## 4.1 Alternative Perspectives

A preliminary testing of the method was carried out with a sample of 30 most frequent nouns in English retrieved from British National Corpus (Kilgarriff, 2009). We found out that when testing the effect of each of five features separately for ranking, relatively different perspectives were achieved for building personalized learning paths.

The gained perspectives can be characterized on

Table 2: Ranking of hyperlinks of article "Life" in respect to five features.

| Rank | Hierarchy of hyperlinks (ordinal number) | | Repetition of hyperlink terms (times) | | Article size (bytes) | | Viewing rate (times per month) | | Editing rate (total number of edits) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Main text* | *Only intro* | *Main text* | *Only intro* | *Main text* | *Only intro* | *Main text* | *Only intro* | *Main text* | *Only intro* |
| 1 | Biota (ecology) 1 | Biota (ecology) 1 | Organism 59 | Organism 59 | Evolution 525544 | Fungus 488952 | Earth 372525 | Earth 372525 | Evolution 12233 | Earth 9152 |
| 2 | Object (philosophy) 2 | Object (philosophy) 2 | RNA 41 | Gene 38 | Fungi 489093 | Metabolism 456427 | Water 286508 | Water 286508 | Earth 9152 | Philosophy 6905 |
| 3 | Biological process 3 | Biological process 3 | Gene 38 | Earth 33 | Fungus 488952 | Earth 417499 | Evolution 206918 | Religion 192527 | Aristotle 7089 | Death 6467 |
| 4 | Death 4 | Death 4 | Earth 33 | Biology 26 | Metabolism 456427 | Bacteria 407412 | Religion 192527 | Philosophy 180609 | Philosophy 6905 | Religion 5850 |
| 5 | Biology 5 | Biology 5 | Evolution 32 | Animal 25 | Bird 440284 | Archaea 354696 | Aristotle 190096 | Animal 173059 | Death 6467 | Water 5828 |
| 6 | Organism 6 | Organism 6 | Biology 26 | Plant 21 | Earth 417499 | Philosophy 220220 | Virus 189972 | Bacteria 153442 | Religion 5850 | Biology 5340 |

various semantic levels, such as causality, conjunction, temporality, quality, space and participants. We think that there is a whole new research domain opening in this ranking-based exploration of wiki environments. Since many statistical uncertainties can bring bias to generalisation if testing does not cover very large experimental samples we want to be cautious when reporting our first preliminary results achieved with relatively small samples. However, to verify overall functionality of our computational method we feel it to be necessary to first start with smaller samples.

To illustrate the rich varied perspectives gained with our method Table 2 shows target articles of high-ranking hyperlinks of Wikipedia article “Life” (October 2009). In each major column of the table hyperlinks are ranked based on each feature separately. The columns “Main text” and “Only Intro” indicate if the ranking is done for all hyperlinks of the full article text or only for hyperlinks mentioned in the introduction section before the table of contents. Applying ranking only to introduction section seems to help highlighting fundamental relations and improves computational performance thus decreasing delay of getting results with the method.

We can sum all types of rankings together for each hyperlink. Then three highest-ranking hyperlinks in descending order for the main text are Evolution, Earth and Organism, and for only the introduction section Earth, Philosophy and Organism. When ranking is done only for the hyperlinks mentioned in the introduction section, the promoted hyperlinks appear to be more sheared among various perspectives than when ranking is done for all hyperlinks of the full article text. In addition, hyperlinks that do not belong to the introduction section and thus are ignored from the ranking based on the introduction section can take high-ranking positions when ranking is based on the full article text. This phenomenon happens with hyperlinks Evolution and Aristotle.

We next evaluate characteristics emerging with each feature in Table 2 against our hypothesis about favourable cases that we outlined in Table 1. In the following we make some brief notions from Table 2. “Hierarchy of hyperlinks” promotes relatively definitive and general hyperlinks. Order of hyperlinks is the same for main text and introduction section, thus no differences in ranking. “Repetition of hyperlink terms” promotes for example gene and RNA that are short words and small structural units. “Article size” promotes hyperlinks to articles that appear to represent major themes and relatively biologically oriented components in life. ”Viewing rate” promotes hyperlinks to rather diverse set of everyday concepts belonging to life. “Editing rate” promotes hyperlinks to articles dealing with actively debated topics such as Evolution, Philosophy, Death and Religion. Overall, we can conclude that these findings match well with our previously made hypothesis about distinctive characteristics for each feature used in ranking of hyperlinks.

## 4.2 Personalized Learning Paths

To evaluate effect of the proposed method to produce personalized learning paths we decided to explore hyperlink chains starting from a Wikipedia article with alternative perspectives. We produced the learning paths in the form of concept maps by exploring hyperlink chains following the ranking in respect to five features described in Section 3.1. We continued testing with the previous sample of 30 English nouns. Due to convincing but diverse outcome we suggest extensive further evaluation with large samples. We introduce some preliminary results illustrated with the case of Wikipedia article “Life”.

Figure 1 shows concept maps that we produced for each of five . perspectives when taking into

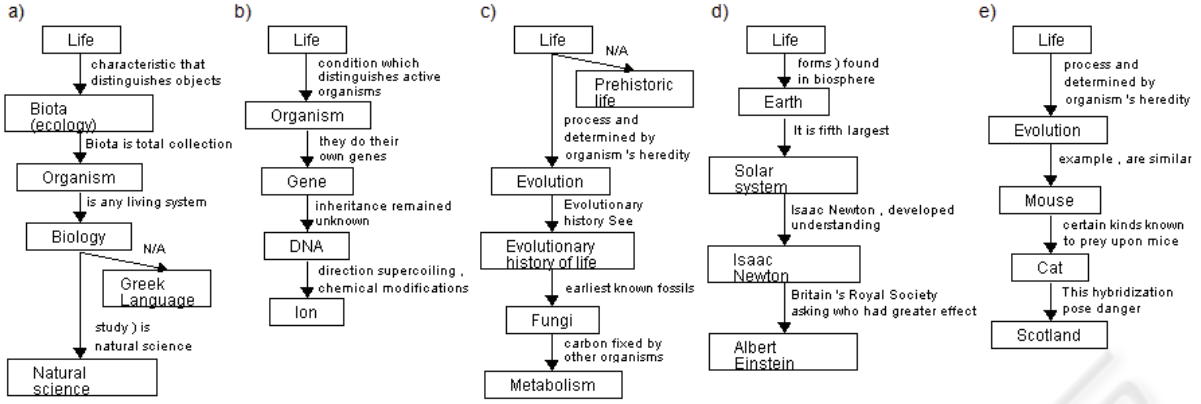


Figure 1: Learning paths starting from article "Life" with five alternative perspectives: a) hierarchy of hyperlinks, b) repetition of hyperlink terms, c) article size, d) viewing rate, and e) editing rate.

account all hyperlinks in the full article text. Relation statements were extracted from sentences surrounding the hyperlinks with a method introduced in our previous work (Lahti, 2009). Due to space constraints, we show here just short linear versions of concept maps that rely on traversing only the highest-ranking hyperlink at each expanding step. However, the method enables the learner to build concept maps with a free design in respect to branching, interconnections and loops. When generating Figure 1, if the highest-ranking hyperlink pointed to an already visited article, a hyperlink with lower rank was accepted to avoid looping. Same was done when relation statement extraction failed to produce a phrase, due to two special segments of the article. Evaluating the learning paths introduced by each of the five features gives promising results. The learning paths seemed to offer some distinctive perspectives corresponding to our previous hypothesis that we outlined in Table 1.

"Hierarchy of hyperlinks" (Figure 1a) produced a learning path that remains constantly on relatively high level of conceptual hierarchy in the topic. This type of learning paths could effectively introduce for example main chapters of the curriculum. "Repetition of hyperlink terms" (Figure 1b) produced a path that goes through conceptual structures of the topic across various hierarchical levels. This type of path could suit well to learning how the curriculum in deeper levels relies on rich variations of some basic components. "Article size" (Figure 1c) produced a path highlighting a collection of the most broadly documented concepts of the topic. This type of path could help in getting idea about the most respected and stabilized parts of the curriculum. "Viewing rate" (Figure 1d) produced a path showing those concepts of the topic that get the most attention from the general public. This type of path could indicate which parts of the curriculum are the most referenced ones. "Editing rate" (Figure 1e) produced a path that offers concepts in the topic that are actively debated by the general public. This type of path could illustrate the parts of curriculum that are involved in constructive criticism and reconsideration.

We think that these preliminary results give support for the proposed method and motivate further research. We think that altering the use of different rankings provided by the five features enables the learner to explore the article network of Wikipedia adaptively. Following his/her intuition and consideration the learner can address current pedagogical needs by selecting the most suitable type of ranking and then the most promising hyperlink. Thus, the learner himself/herself can build personalized learning paths for various educational purposes. Building learning paths is naturally very sensitive for each consecutive step of selecting the most promising hyperlink in the current article for further exploration. Additional testing indicated that selecting the hyperlinks only from the introduction section of the article increased keeping the steps of the learning path more closely related. We think that the learner needs to make several parallel branching learning paths to ensure sufficient diverse coverage of the learning topic.

# 5 RELATED WORK

## 5.1 Ontology Construction

To bring structure to the meaningful content of web pages, so called semantic web approach aims to introduce *ontologies* as a formal representation for concepts within a domain and the relationships

between them (Berners-Lee et al., 2001). However, many traditional ontology projects have received criticism about being too closed, formal and hard to update (Simperl & Tempich, 2006). Zouaq and Nkambou (2009) proposed a method to automatically generate a domain ontology from plain text documents and use this ontology as the domain model in computer-based education. They suggested evaluating the generated domain ontology with three dimensions: structural, semantic, and comparative. Pirrone et al. (2005) proposed an approach to automated learning path generation inside a domain ontology supporting a web tutoring system. Inspired by the knowledge space theory, they suggest heuristics to transform an ontology in a weighted graph where the A* search algorithm is used to find the path.

We think that a community-driven approach, such as wiki environments, can well support dynamic collaboratively defined ontologies. The Wikipedia does not have permanently fixed categorization of its content and the relations can sustain even radical changes to respond the changes in the average worldview. The content providers are asked to take care of updating the organization of the content as well. Since previous versions can be always reverted, it is safe to let the structure freely slowly converge towards a consensus while complementary contributions are gathered.

Despite uncertainties, the Wikipedia has been considered as a promising source for ontology construction (Haase & Völker, 2008; Hu, in press). Every Wikipedia article describes one concept denoted by the title of the article that has been considered having value for general public. Each hyperlink of this article literally shows a path to another related concept that has been collectively valued so much that a specific article has been written about it as well.

With swarm intelligence, spontaneous indirect coordination between agents can show optimal learning paths with a form of self-organization called stigmergy (Gutiérrez et al., 2006). Similarly, we think that automated generation of favourable learning paths can be effectively based on proceeding in the conceptual network represented by Wikipedia articles and inter-article hyperlinks.

Graph based visualizations relying on ontologies extracted from the Wikipedia have been proposed for education (Dicheva & Dichev, 2007; Yang et al., 2007). We now suggest extending the use of ontologies extracted from the Wikipedia to be applied in building personalized learning paths. This poses requirements to assess the quality of articles and perspectives that they can provide.

With an aim to enhance the quality of articles, the Wikipedia community has been labelling in a specific review process some satisfactory articles as "good articles" and even more professional ones as "featured articles". Blumenstock (2008) showed that the featured articles can be recognized correctly with the accuracy of 96 % using a simple heuristic that classifies articles with more than 2000 words as "featured" and articles with fewer than 2000 words as "random". Thomas and Sheth (2007) showed that when comparing labelled good articles to other non-stub articles having at least 50 revision milestones they found no statistically significant difference in convergence to a semantically stable state.

These two previous results indicate that the maturity of an article can be measured relatively well even with simple parameters. This seems to support our attempt to identify few basic features of an article that can be easily measured to create rankings for hyperlinks, highlighting alternative perspectives that they provide. Adoption of features denoting "Article size" and "Editing rate" for our method was especially inspired by these two previous results.

Since each collaborating author of a Wikipedia article inherently provides a different perspective, the authors can together easily cover a rather wide range of issues and typically produce a rich creative output. The iterative additions, evaluations and edits done collectively seem to be capable of assuring the quality and coherence of the steps leading towards the desired stabilized state of a mature article.

## 5.2 Accumulating Knowledge

Semantic relation analysis tries to identify semantic units and their relations from natural language. Previous research has strongly relied on machine learning approach or probabilistic methods. Semantic roles have been identified for example from syntactic units by using various lexical resources and predefined relations. When trying to generate automatically favourable learning paths that match with the learner's needs the guidance should have a suitable balance between constraints for sustainability and freedom of association.

Nastase and Szpakowicz (2006) introduced an incremental learning algorithm that effectively mimics the way in which a human reader accumulates knowledge and exploits it to process new text. The algorithm assigns a semantic relation to semantic units of text taken from a science book with guidance from a user and builds a simple syntactic-semantic graph surrounding the central concept and matching it with previously analysed

text. Our proposed method provides an analogous approach relying on user-driven generation of learning paths in the form of concept maps based on article network of the Wikipedia.

Coursey et al. (2008) argued that a combination of keyword extraction techniques combining graph-theoretical algorithms and methods relying on knowledge extracted from the Wikipedia can be successfully used to identify candidate keywords in learning objects. They suggest using ranking algorithm over the Wikipedia connectivity graph to find relevant articles. Somewhat similarly, our method exploits the titles of hyperlink's target articles to identify promising concepts in the Wikipedia. We introduce ranking that enables these concepts to be explored in learning paths, accompanied with compact relation statements parsed from the sentences surrounding each hyperlink.

Serrano et al. (2009) argued that some key regularities of written text concerning burstiness of words, topicality and their relationship can be modelled with two simple algorithmic techniques that are frequency ranking with dynamic reordering and memory effect connecting word frequencies across different documents. They suggest that their model enables to relate two key mechanisms that have been assumed to affect how humans process the lexicon: rank frequency and context diversity. They propose using their model to study co-evolution of content and citation structure for example in the Wikipedia. In a resembling fashion, our method uses rank frequency and context diversity of the Wikipedia enabling a learner to process lexicon to a pedagogically rewarding structure.

There have been proposals to visually highlight the most mutually agreed segments in a Wikipedia article based on simple quality measures. High survival time of a single edit does not guarantee reliably its trustworthiness (Luyt et al., 2008). Adler and de Alfaro (2007) proposed a measure relying on author's reputation that can be gained if the edits he/she performs are preserved by subsequent authors. It seems challenging to develop measures taking simultaneously into account the semantics of the article network and collective contribution patterns coming from authors and readers. Our method tries to address these issues.

Pavlovic (2008) introduced a model of network computation based on Markov chains as an attempt to extract the semantic content of the data from their distribution among the nodes. A concept can be identified by finding the community of nodes that share it performing together some non-local data processing. Pavlovic proposes ranking of paths to extract information about the likely flow biases from the available static information about the network and thus to detect semantic structures in a network. We think that our method is dealing with a same kind of goal and that the statistics concerning articles can be useful criteria for ranking the paths.

Haruechaiyasak and Damrongrat (2008) proposed a method to generate a topic model from articles in the Wikipedia Selection for Schools that is a collection aiming to meet curricula world-wide. Their method relies on latent Dirichlet allocation. Based on similarity measures computed for topic distribution profiles of the articles, the method enables recommending related articles some of which are not covered by the hyperlinks in a given article. With some similarities, our method aims to recommend well related hyperlinks for learning, We decided to use article statistics as a measure that enables to highlight different perspectives..

All in all, inspired by the previous work, we tried to find measures to recommend hyperlinks that offer pedagogically motivated exploration in the Wikipedia article network with a preferred perspective. To make learning paths personalized there is a need for a method that takes into account the learner's preference about which hyperlink is the most profitable to choose as the next step in the learning path. To find a favourable chain of hyperlinks in the Wikipedia a reasonable amount of familiarity and continuity should be preserved while still trying to extend knowledge of the learner.

# 6 DISCUSSION

The proposed method aims to suggest hyperlink chains that offer highest pedagogic value for the learner. An essential strength of the method is the aim to provide a reasonable collection of alternative hyperlink chains that maintain semantic and educational relatedness between each step in the chain and between parallel chains. This is based on four key factors: collaboratively maintained initial organization of concepts and relations (evolution of the Wikipedia), dynamic ranking in respect to five features supporting alternative perspectives (article statistics), illustrations denoting previous and current conceptual reasoning (concept maps), and letting the learner to make the ultimate decision for next step based on her intuition and consideration (support for variety of personalities).

The proposed method relies heavily on extraction and analysis of hyperlinks in Wikipedia articles related to a chosen learning topic. Recommendable

learning paths are represented as a gradually expanding concept map that can be directly shown to the learner and also applied later in various educational purposes.

The method aims to provide a balanced trade-off between extensive coverage and compactness in the generated learning content. The method offers learning paths that should enable the learner to traverse the most essential knowledge in the least amount of time. This traversing can be exploited as means to adopt new knowledge or to refresh it. The traversed learning paths become documented as concept maps thus enabling the learner to analyse illustratively her conceptualization concerning a chosen topic. These knowledge structures can be easily further edited, reused and shared with other learners.

In contrast with many previous proposals in this field, we have not only developed a method for knowledge management but also implemented a fully functional prototype that is ready to be used in various educational contexts for many pedagogical purposes. We do not know any previous proposal that is similar to our work especially concerning educational use of the Wikipedia.

We have evaluated ranking hyperlinks of the article in respect to five different features based on article statistics. We think that these features correspond to the most fundamental functions of the Wikipedia. In our experiments we found distinctive ways to differentiate exploration of hyperlinks based on the features preferred by the learner. Using various rankings it is thus possible to provide alternative perspectives to knowledge and thus enable the learners to build independently favourable learning paths following their personal needs at the moment.

Concepts belonging to various domains of life and to various abstraction levels in a certain topic have obviously different tendencies to support the five features. Also, features can have many hidden correlations that should be taken into account for a balanced use of statistics. High editing rate typically produces high article size. Typically each single event of editing article increases also viewing rate if the editor wants to check the finished version. When building learning paths, our proposed method possibly too optimistically expects high relation between all consecutive concepts in a traversed chain of hyperlinks.

Hyperlinks of an article can point to target articles that deal with topics that are opposite or ambiguous to the title of current article. Unfortunately, it is hard to develop general methods that could reliably identify the exact type of relation between target article and current article. Extracting relation statements from the sentence surrounding the hyperlink can also be troublesome since often the sentence does not explicitly define the relationship between the title of current article and the title of target article, but instead describe something else.

When building learning paths, a major challenge for semantic continuity is that some measures based on the characteristics of target article may not indicate well the actual relatedness between current article and target article. For example, if ranking of hyperlinks is based on viewing rate, the target article having the highest viewing rate is prioritized. But this viewing rate consists of a great variety of visits arriving to the target article through various hyperlinks, not only from current article. Thus, viewing rates describe just the overall distribution of visits to individual Wikipedia articles and fail to tell how the preference to visit a certain target article varies depending on the current article.

There are limitations with the current method especially since it was purposefully designed to be simple and computationally easy. The features used with the method could be chosen in various alternative ways. Anyway, if the online services used for quering statistics should become shut down it still remains possible to retrieve statistics with alternative implementations. The method might be enhanced by using statistics taken from varied time frames and making diverse temporal analysis for views, edits, article size, etc. Articles could be treated more equally if comparison would rely on proportional values instead of absolute values. Thus, view rates and edit rates could be considered proportionally, for example in relation to article size. The method could also somehow take advantage of the fact that typically many changes in an article are performed in bursts, for example after related news has been published in the media.

Various navigational aids have been introduced to the layout of Wikipedia articles, for example category tags, “See also” section, naviboxes and infoboxes. Also redirects, disambiguation pages and “What links here” queries assist finding related articles. However, we argue that these assistive functions complementing each other cannot clearly recommend the most promising hyperlinks for further exploration. To increase efficiency of exploration and to ensure finding the most relevant hyperlinks, there is a need for adaptive ranking of hyperlinks of the current article.

Since a lot of articles of the Wikipedia present facts that have a low probability to become constantly updated or seriously questioned, we think

that our method could be successfully used also off-line. Despite of its huge coverage, the plain textual content of English language version of Wikipedia can be stored locally in one compressed file of 5 gigabytes. The method might use also the article statistics from just off-line sources. We suggest that already the statistics gathered so far can enable creating reliable ranking that reflects conception of global community. Relying on off-line content would enable using the method with very low computational costs and minimal delay.

# 7 FUTURE WORK

The proposed method and experiment have indicated a promising unexplored area for research concerning new methodology to adaptively explore the knowledge space of the Wikipedia. We suggest that the method we have developed for the Wikipedia can be relatively well applied to also other collaborative knowledge management environments and even intellectual mental processes in human mind.

In the future research there is a strong need for further classifying various features that can be used in ranking of hyperlinks that connect concepts (or articles). It can be possible to identify individual most favourable features for each domain of knowledge. These specific features could enable exploring knowledge in most coherent manner taking into account special characteristics that are typical for this domain.

It is also important to develop methods that can address individual characteristics of every learner. For each learner it could be identified what are the features that need to be used in ranking of hyperlinks to fulfil his/her special personal needs. For example, preferred learning style, personality and hobbies of the learner could be considered when setting the ranking criteria that affect which hyperlinks become promoted to the learner. Furthermore, it would be advantageous that the learner could himself/herself make adaptively consistent decisions about what features to prioritize in ranking when exploring varying learning contexts. In many cases, user-defined ranking criteria should not probably support just one perspective but instead to be a dynamic weighted mixture of them all.

In addition, it is important to develop advanced but still computationally sustainable analysis methods that help to rank alternative hyperlinks and thus to find most promising learning paths. It is important to have such analysis methods that are not dependent on any proprietary online service. To effectively develop and ensure automated knowledge management it is important to support open access knowledge bases and open source modules. Interfaces should be kept as interoperable and standardized as possible to best promote updating individual components of modular applications or replacing them with alternatives.

Knowledge management tools should be actively introduced for using them in ordinary life for example in education, problem solving, decision making, design and innovation. Research should emphasize access for all since knowledge tools are often most crucial for people with special needs. The efforts should aim at providing a better quality of life and letting the learner to excel oneself and follow his/her personal interests.

# REFERENCES

Adler, B., & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. Proc. 16th international conference on World Wide Web, Banff, Alberta, Canada, ACM Press, 261-270.

Berners-Lee, T., Hendler, J. & Lassila; O. (2001). The semantic web. Scientific American Magazine, May 2001.

Blumenstock, J. (2008). Automatically assessing the quality of Wikipedia articles. University of California at Berkeley, School of Information. Technical Report 2008-021.

Chesney, T. (2006). An empirical examination of Wikipedia's credibility. First Monday, 11(11).

Coursey, K., Mihalcea, R., & Moen, W. (2008). Automatic keyword extraction for learning object repositories. Proc. Conference of the American Society for Information Science and Technology (ASIST 2008), Columbus, Ohio, USA.

Dicheva D., & Dichev C. (2007). Helping courseware authors to build ontologies: the case of TM4L. Proc. 13th International Conference on Artificial Intelligence in Education, (AI-ED 2007), Los Angeles, California, USA, IOS Press, 77-84.

Gandrabur, S., Foster, G., & Lapalme, G. (2006). Confidence estimation for NLP applications. Transactions on Speech and Language Processing, 3(3), 1-29.

Guthrie, J., Wigfield, A., Barbosa, P., Perencevich, K., Taboada, A., Davis, M., Scafiddi, N., & Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. Journal of Educational Psychology, 96(3), 403-423.

Gutiérrez, S., Pardo, A., & Kloos, C. (2006). Finding a learning path: toward a swarm intelligence approach, Proc. 5th IASTED international conference on Web-based education, Puerto Vallarta, Mexico, ACTA Press, 94-99.

Haase, P., & Völker, J. (2008). Ontology learning and reasoning – dealing with uncertainty and inconsistency. In da Costa, P. et al. (eds.), Uncertainty Reasoning for the Semantic Web I. LNAI 5327, 366-384.

Haruechaiyasak, C., & Damrongrat, C. (2008). Article recommendation based on a topic model for Wikipedia Selection for Schools. Proc. 11th International Conference on Asian Digital Libraries, LNCS 5362, 339-342.

Holmes, B., Tangney, B., Fitz-Gibbon, A., Savage, T., & Mehan, S. (2001). Communal constructivism: students constructing learning for as well as with others. Proc. 12th International Conference of the Society for Information Technology and Teacher Education (SITE 2001), Orlando, Florida, USA, 3114-3119.

Hu, B. (in press). WiKi'mantics: interpreting ontologies with Wikipedia. Journal of Knowledge and Information Systems, Springer. DOI: 10.1007/s10115-009-0247-6

Janssen, J., Berlanga, A., Vogten, H., & Koper, R. (2008). Towards a learning path specification. International Journal of Continuing Engineering Education and Lifelong Learning, 18(1).

Kilgarriff, A. (2009). BNC database and word frequency lists. Lemmatized frequency list of British National Corpus. URL: http://www.kilgarriff.co.uk/bnc-readme.html

Lahti, L. (2009). Guided generation of pedagogical concept maps from the Wikipedia. Proc. World Conference on E-Learning in Corporate, Government, Healthcare and Higher Education (E-Learn 2009), Vancouver, Canada, 1741-1750. ölvlably pwjxjyt åjtdtj, pwzyt ldltybcvtulb jtbt taj bccwt, vttbza.

Luyt, B., Aaron, T., Thian, L., & Hong, C. (2008). Improving Wikipedia's accuracy: Is edit age a solution? Journal of the American Society for Information Science and Technology, 59(2), 318-330.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In Schmitt, N., & McCarthy, M. (eds.), Vocabulary: Description, acquisition, pedagogy. Cambridge University Press, New York, USA, 6-19.

Nastase, V., & Szpakowicz, S. (2006). Matching semantic-syntactic graphs for semantic relation assignment. Proc. Textgraphs 2006 Workshop on Graph-based Algorithms for Natural Language Processing, New York, USA.

Neumann, D., & Hood, M. (2009). The effects of using a wiki on student engagement and learning of report writing skills in a university statistics course. Australasian Journal of Educational Technology, 25(3), 382-398.

Pavlovic, D. (2008). Network as a computer: ranking paths to find flows. Proc. Third International Computer Science Symposium in Russia, LNCS 5010, 384-397.

Peregrin, J. (in press). The myth of semantic structure. In Stalmaszczyk, P. (ed.), Philosophy of Language and Linguistics. Volume I: The Formal Turn. Ontos, Frankfurt. Germany. URL: http://jarda.peregrin.cz/mybibl/PDFTxt/528.pdf

Pirrone, R., Pilato, G., Rizzo, R., & Russo, G. (2005). Learning path generation by domain ontology transformation. Proc. 9th Congress of the Italian Association for Artificial Intelligence, LNAI 3673, 359-369.

Reinoso, A., Ortega, F., Gonzalez-Barahona, J., & Robles, G. (2009). A quantitative approach to the use of the Wikipedia. Proc. IEEE Symposium on Computers and Communications (ISCC 2009), Sousse, Tunisia, 56-61.

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. Psychological Science, 3(4), 207-217.

Serrano, M., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. Public Library of Science ONE (PLoS ONE), 4(4): e5372.

Simperl, E., & Tempich, C. (2006). Ontology engineering: a reality check. Proc. 5th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE2006), LNCS 4275, 836-854.

Thomas, C., & Sheth, A. (2007). Semantic convergence of Wikipedia articles. Proc. IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, California, USA, 600-606.

Van Berkum, J., Brown, C., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. Journal of Experimental Psychology: Learning, Memory and Cognition, 31, 443-467.

Wikipedia article traffic statistics (2009). URL: http://stats.grok.se/

Wikipedia page history statistics (2009). URL: http://vs.aka-online.de/cgi-bin/wppagehiststat.pl

Yang, J., Jangwhan, H., Oh, I., & Kwak, M. (2007). Using Wikipedia technology for topic maps design. Proc. 45th ACM Southeast Regional Conference (ACM-SE 45). Winston-Salem, North Carolina, USA, ACM Press, 106-110.

Zouaq, A. & Nkambou, R.,(2009). Evaluating the generation of domain ontologies in the Knowledge Puzzle Project. IEEE Transactions on Knowledge and Data Engineering, 21(11), 1559-1572.