Educational tool based on topology and evolution of hyperlinks in the Wikipedia

Lauri Lahti
Department of Computer Science and Engineering
Aalto University School of Science and Technology (TKK), Finland

***Abstract*—We propose a new method to support educational exploration in the hyperlink network of the Wikipedia online encyclopedia. The learner is provided with alternative parallel ranking lists, each one promoting hyperlinks that represent a different pedagogical perspective to the desired learning topic. The learner can browse the conceptual relations between the latest versions of articles or the conceptual relations belonging to consecutive temporal versions of an article, or a mixture of both approaches. Based on her needs and intuition, the learner explores hyperlink network and meanwhile the method builds automatically concept maps that reflect her conceptualization process and can be used for varied educational purposes. Initial experiments with a prototype tool based on the method indicate enhancement to ordinary learning results and suggest further research.**

Keywords—knowledge maturing; content-based filtering; semantic relatedness; concept map; Wikipedia

I. Introduction

Collaboratively built Wikipedia online encyclopedia (http://en.wikipedia.org), has revolutionized gathering and sharing knowledge with open source movement. The Wikipedia is considered as a scale-free network (a small world network) [1] that automatically evolves to a hierarchical clustering structure following so called power law making new vertices attached preferentially to already well connected nodes [2]. Despite mixed acceptance from educators [3], the coverage and quality of the Wikipedia is said to meet the level of respected encyclopedias [4] and median survival time for vandalism edits is 11 minutes [5].

We think that a large part of curriculum has already been iteratively elaborated in the articles of Wikipedia. Wikipedia has many collaboratively agreed structural characteristics that intuitively support a learner to find personalized learning material at an appropriate level of complexity. We consider that the Wikipedia can adaptively support personalized learning of concepts and their relations. Each article defines a concept denoted by its title and its hyperlinks define relations to other concepts. We propose a new method helping the learner to explore and analyze semantic relations between concepts represented by Wikipedia articles by using adaptive lists and a concept map. Based on the method we have implemented a prototype to serve as an educational tool to support individual and collaborative learning tasks.

II. Method

The learner provides an initial concept about the learning topic for the method and selects either topological or evolutionary exploration mode. In *topological exploration mode*, the learner proceeds in the network of hyperlinks belonging to the latest versions of Wikipedia articles. The hyperlinks are shown in a few parallel ranking lists providing alternative rankings sorted in decreasing order of significance. Based on distinct ranking criteria, each list promotes hyperlinks representing a different pedagogical perspective to the learning topic. Each row in a ranking list shows the concept attainable through the hyperlink (title of target article), relation statement (excerpt from the sentence surrounding the hyperlink in current article) and a statistical measure from current or target article used to create ranking.

From the ranking lists the learner selects a desired amount of concepts that seem promising for her, indicating what perspectives she wants to be prioritized by the method in further exploration. Selected concepts and their relations to previous concepts become illustrated in a progressively expanding concept map. Nodes labeled with the concepts are connected with directed arcs labeled with relation statements respectively. From the concept map the learner selects one concept for the next step in exploration and from now on each ranking list shows hyperlinks for the article corresponding to this selected concept. By repeating this cycle, step by step, new hyperlinks with alternative rankings are constantly recommended by the method thus providing a diversity of exploration paths. Based on her needs and intuition, the learner explores hyperlink network and meanwhile the method builds automatically a concept map that reflects her conceptualization process.

We suggest that ranking of hyperlinks should rely on simple statistics concerning current article and target article. Based on convincing results in our previous work [6], reflecting five main functions identified for the Wikipedia, we decided to use following measurable parameters as ranking criteria for hyperlinks: order of hyperlinks in current article, hyperlinks whose target article's titles are most repeated in current article, size of hyperlink's target article, view rate of hyperlink's target article and edit rate of hyperlink's target article. These measures can be easily retrieved from revision history and online services providing Wikipedia statistics, and relation statements can be extracted from sentences surrounding hyperlinks with a parsing method, as explained in our previous work.

In *evolutionary exploration mode* a concept and its relations can be represented by any previous temporal version of the corresponding Wikipedia article and its hyperlinks at that time. The learner is provided with a simple dial to select a desired time frame from the revision history of current article. Also the ranking of hyperlinks is carried

Lahti, L. (2010). Educational tool based on topology and evolution of hyperlinks in the Wikipedia. Proc. 10th IEEE International Conference on Advanced learning Technologies (ICALT 2010), 5–7 July 2010, Sousse, Tunisia.

out with statistics from the chosen historical moment in time. The learner can browse consecutive temporal versions of articles to see how new hyperlinks and relation statements are introduced and how older ones become edited or even removed. By observing these temporal transformations the learner can get insight how conceptualization can proceed in a collaborative environment. By alternating between both evolutionary and topological exploration modes, the learner should receive even additional pedagogical advantage as she simultaneously gives attention to both temporal local emergence of knowledge clusters and general connectivity among clusters in relations fixed to a certain time frame.

We propose two optional enhancements for the method that are memory effect and definition boost. *Memory effect* gives extra promotion for hyperlinks that are shared among concepts added so far to the concept map. If at least two previously encountered articles have a same target article as the current article has, this hyperlink will be automatically given a leading position in the ranking lists. *Definition boost* lets the learner to see only those hyperlinks belonging to the introduction section of current article, typically before the table of contents. Since writing style in introduction is often more definitive than later in the article, also recommended hyperlinks are expected to emphasize now more definitions.

To demonstrate the method, map a) in Figure 1 shows a concept map generated with the topological exploration mode starting from Wikipedia article “History of the world”. To avoid self-serving bias, a symmetric tree topology was used with uniform ranking criteria based on sum of all five suggested rankings. On each level the sibling nodes are shown from left to right in descending order of ranking. To emphasize initial diversity, memory effect was applied only to generate the nodes of third level. Definition boost was applied on all levels. We think that resulting concept map offers rather relevant exploration paths. Memory effect promoted four useful concepts to be included in the map (Indian subcontinent, Prehistory, Greek language and Early modern period). Maps b), c) and d) in Figure 1 show stubs of concept maps generated with evolutionary exploration mode based on three time frames of article “History of the world” using similar conditions as described for map a). We think that traits of maturing can be seen in this temporal series.

In initial experiments, we compared the conceptual structures generated with our method, to corresponding established learning material. For example, Figure 1 matched well with exploration paths gained when accessing four main periods of history through index of a children’s world history book [7]. Initial test users reported that method fruitfully enhanced their ordinary learning results, supplied with representations successfully balancing between simplicity and accuracy that can be easily complemented and finetuned to reach an even further educational gain.

III. Related work

Educational tools providing holistic solutions for everchanging learning scenarios are needed [8]. As an intuitive medium, concept maps have been recommended for illustrating relationships of educational material in both flexible and compact form [9]. Knowledge maturing has been verified in the Wikipedia as implicit contextualized knowledge becomes gradually explicitly linked and formalized, and useful measures for maturing can possibly be extracted from creation and usage contexts [10]. To exploit the maturing of Wikipedia for pedagogical exploration, our work is inspired by intelligent tutoring systems, content-based filtering, information retrieval and clustering. Weber et al. [11] introduced a tool for visual semantic browsing and decision making based on concept maps. García-Plaza et al. [12] proposed an unsupervised document representation model to cluster web pages with

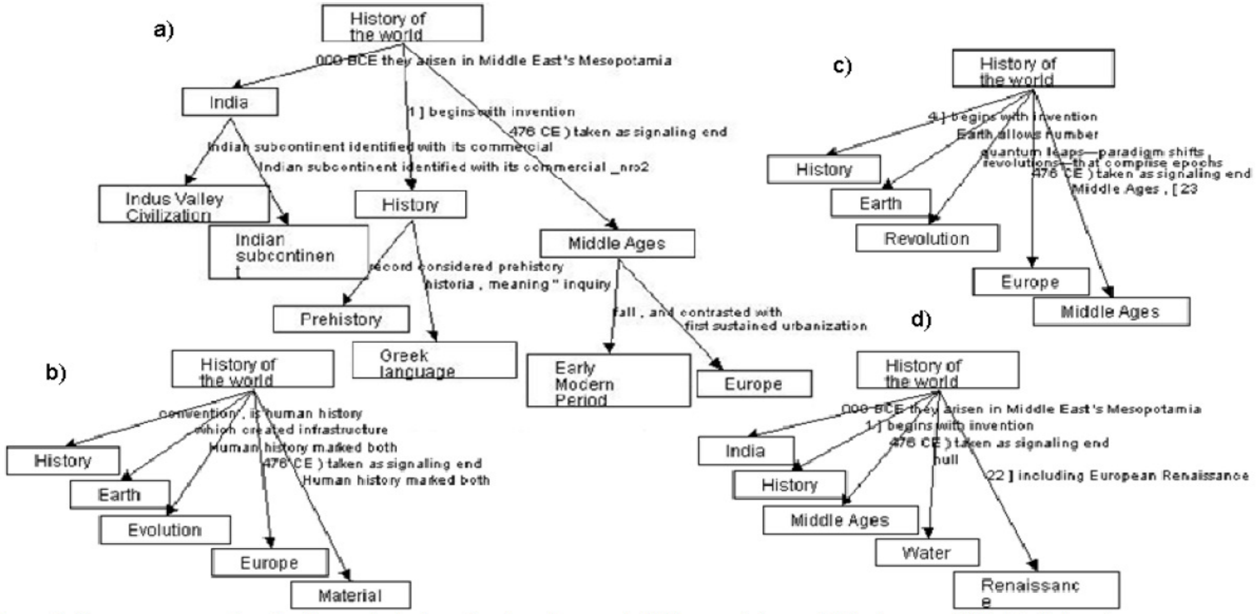
a)
History of the world
000 BCE they arisen in Middle East's Mesopotamia
1] begins with invention
476 CE) taken as signaling end
India
Indian subcontinent identified with its commercial
Indian subcontinent identified with its commercial _nro2
Indus Valley Civilization
Indian subcontinent
History
Middle Ages
record considered prehistory
historia , meaning " inquiry
Prehistory
Greek language
fall , and contrasted with
first sustained urbanization
Early Modern Period
Europe
b)
History of the world
convention", is human history
which created infrastructure
Human history marked both
476 CE) taken as signalling end
Human history marked both
History
Earth
Evolution
Europe
Material
c)
History of the world
4] begins with invention
Earth allows number
quantum leaps—paradigm shifts
revolutions—that comprise epochs
476 CE) taken as signaling end
Middle Ages , [23
History
Earth
Revolution
Europe
Middle Ages
d)
History of the world
000 BCE they arisen in Middle East's Mesopotamia
1] begins with invention
476 CE) taken as signaling end
null
22] including European Renaissance
India
History
Middle Ages
Water
Renaissance

Figure 1. Concept map produced with topological exploration about topic ”History of the world” in January 2010 (a). Stubs of concept maps produced with evolutionary exploration about topic ”History of the world” with three time frames: January 2008 (b), January 2009 (c), and January 2010 (d).

self-organizing maps using features of the pages. These works support us to develop map-based tool for exploration without extensive indexing of the Web. Hyperlinks can be seen as a tagging about the article's context. Kamps and Koolen [13] showed that the degree of arriving hyperlinks can be exploited to significantly improve effectiveness of ad hoc information retrieval. Zubiaga et al. [14] showed that socially annotated web content can be well classified based on weighted tags, even with limited user counts. Noll and Meinel [15] showed that tag-based classification seem to suit better to top-level documents in a hierarchy and deeper levels need contextual information mediated from higher levels. These results motivate us to recommend hyperlinks for exploration based on simple ranked statistics about articles that are hierarchically related or encountered earlier. To address imprecision, Kotsakis [16] proposed querying XML documents with fuzzy ranking relying on Levenshtein distances based on tags encountered in paths and characters included in terms. Emphasizing document's structure, Cafarella et al. [17] proposed querying relational information from HTML tables on the Web and ranking them in respect to diverse text-derived features. To integrate schema information from numerous structured data sources, Nandi and Bernstein [18] proposed a semi-supervised mapping method relying on a log of queries that cause click-throughs. The DBpedia [19] is a promising project extracting structured factual information from Wikipedia articles to form an expressive dataset facilitating queries about relationships and properties. Chan et al. [20] proposed a search algorithm over the DBpedia enabling to extract a semantic graph from Wikipedia's hyperlink structure. Another interesting effort to exploit the Wikipedia is semantic search engine NAGA [21] using graph-based query language with ranking that considers confidence, informativeness and compactness of results.

IV. Discussion and future work

Even a short chain of hyperlinks in the Wikipedia can cover essential knowledge about a desired educational topic. Due to rich variety of contributors, the hyperlink network of the Wikipedia combines numerous individually favored relations between concepts into one browsable entity. However, it is hard to define requirements for optimal exploration paths that can be favorably personalized in diverse contexts and generated with limited computational load.

Results of related work indicate that simple quantitative semi-automatic methods can be successfully used for measuring matching with imprecise queries to rank documents in a collection. This suggests that desired educational perspectives can be efficiently promoted by chaining ranked hyperlinks that have even relatively imprecise correlation between a simple statistical feature of current and target article. To enable holistic adaptive conceptualization process, the learner needs interactive knowledge representations and concept maps seem to offer an efficient medium for compact yet flexible illustrations. Besides exploring just the relations between the latest versions of articles, browsing consecutive temporal versions of an article enables analyzing emergence of knowledge clusters. Two additional options enable to favor hyperlinks having previously encountered target articles and hyperlinks promoting definitions. Initial experiments with a prototype indicate that proposed functional principles can fruitfully support exploration that is sustainable for human learning.

Future work should address agglomeration of separate learning tasks and complementing collaboration schemes. Easy evaluation and intervention methods are needed for teachers. Personal learning styles and special needs should be strongly supported with encouragement and inspiration.

References

[1] Capocci, A., Rao, F., & Caldarelli, G. (2008). Taxonomy and clustering in collaborative systems: the case of the on-line encyclopedia Wikipedia. Europhysics Letters, 81(2).

[2] Barabási, A., & Réka, A. (1999). Emergence of scaling in random networks. Science, 286:509-512.

[3] Watson, K., & Harper, C. (2008). Supporting knowledge creation: using wikis for group collaboration. Educause Center for Applied Research, Research Bulletin, Issue 3, 2008.

[4] Giles, G. (2005). Internet encyclopaedias go head to head. Nature, 438, 7070, 900-901.

[5] Kittur, A., Suh, B., Pendleton, B., & Chi, E. (2007a). He says, she says: conflict and coordination in Wikipedia. Proc. CHI 2007.

[6] Lahti, L. (2010). Personalized learning paths based on Wikipedia article statistics. Proc. CSEDU 2010. L<3P

[7] Adams, S. (2008). Children's atlas of world history. Kingfisher, Macmillan Children's Books, Singapore.

[8] Utz, W., Hrgovcic, V., & Karagiannis, D. (2009). ADVISOR: towards holistic model-based e-learning environments based on metamodelling concepts. Proc. m-ICTE 2009.

[9] Buzan, T., & Buzan B. (2003). The mind map book. BBC Worldwide Limited, London.

[10] Braun, S., & Schmidt, A. (2007). Wikis as a technology fostering knowledge maturing: what we can learn from Wikipedia. Proc. IKNOW 2007.

[11] Weber, N, Schoefegger, K., Bimrose, J., Ley, T., Lindstaedt, S., Brown, A., & Barnes, S. (2009). Knowledge maturing in the Semantic MediaWiki: a design study in career guidance. Proc. ECTEL 2009.

[12] García-Plaza, A., Fresno, V., & Martínez, R. (2008). Web page clustering using a fuzzy logic based representation and selforganizing maps. Proc. WI-IAT 2008.

[13] Kamps, J., & Koolen, M. (2008). The importance of link evidence in Wikipedia. Proc. ECIR 2008.

[14] Zubiaga, A., Martínez, R., Fresno, V. (2009). Getting the most out of social annotations for web page classification. Proc. DocEng 2009.

[15] Noll, M. & Meinel, C. (2008). The metadata triumvirate: social annotations, anchor texts and search queries. Proc. WI-IAT 2008.

[16] Kotsakis, E. (2006). XML Fuzzy Ranking. Proc. FQAS 2006.

[17] Cafarella, M., Halevy, A., Wang, Z., Wu, E., & Zhang, Y. (2008). Webtables: exploring the power of tables on the web. Proc. VLDBE 2008.

[18] Nandi, A., & Bernstein, P. (2009). HAMSTER: using search clicklogs for schema and taxonomy matching. Proc. VLDBE 2009.

[19] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. Journal of Web Semantics, 7(3), 154-165.

[20] Chan, B., Wu, L., Talbot, J., Cammarano, M., & Hanrahan, P. (2008). Vispedia: interactive visual exploration of Wikipedia data via searchbased integration. Proc. IEEE Information Visualization 2008.

[21] Kasneci, G., Suchanek, F., Ifrim, G., Ramanath, M., & Weikum, G. (2008). NAGA: searching and ranking knowledge. Proc. ICDE 2008.