Author(s): Lahti, Lauri

Title: Educational exploration based on conceptual networks generated by students and Wikipedia linkage

Year: 2014

Version: Post print

Please cite the original version:

Lahti, Lauri. 2014. Educational exploration based on conceptual networks generated by students and Wikipedia linkage. World Conference on Educational Multimedia, Hypermedia and Telecommunications 2014 (EdMedia 2014), Tampere, Finland, 23-27 June 2014. P. 964-974. ISBN 978-1-939797-08-7 (electronic).

Lahti, L. (2014a). Educational exploration based on conceptual networks generated by students and Wikipedia linkage. Proc. World Conference on Educational Multimedia, Hypermedia and Telecommunications 2014 (EdMedia 2014). 23–27 June 2014, Tampere, Finland (eds. Herrington, J. et al.), 964–974. ISBN 978-1-939797-08-7. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA, USA. http://www.editlib.org/p/147608/

Educational exploration based on conceptual networks generated by students and Wikipedia linkage

Lauri Lahti
Department of Computer Science and Engineering
Aalto University School of Science, Finland
lauri.lahti@aalto.fi

Abstract: We propose a new educational framework for educational exploration based on conceptual networks generated and explored by students supplied with Wikipedia linkage. In the first experimental setup we had a group of students (n=103) to create high-frequency lists of words and links between words, and in the second experimental setup we had another group of students (n=49) to explore a subsection of hyperlink network of the Wikipedia corresponding to conceptual networks generated by students in the first experimental setup. We report findings based on comparison of word lists and conceptual networks generated by students, vocabulary ranking of British National Corpus, hyperlink network structure of the Wikipedia and exploration paths of students in the hyperlink network of the Wikipedia. After traversing 20 hyperlink steps each student could recall on average about 33,1 percent of unique shown concepts and on average 64,8 percent of unique selected concepts which seems to indicate that exploration in hyperlink network can support adoption of new knowledge.

Introduction

We think that it is important to find answers to persistent challenge of generating guidance for personalized exploration in knowledge structures and supporting agglomerating and linking pieces of knowledge in a pedagogically fruitful way. Our proposals are inspired by adaptive and efficient link structures that have properties of so called small-world networks and scale-free networks and even both of them together.

Small-world topology has been identified structurally and functionally in human brain networks (Wang et al. 2010) and also scale-free properties have been possibly identified in human brain networks (Bullmore & Sporns 2009). Furthermore, one of the currently biggest open knowledge resources Wikipedia online encyclopedia (http://www.wikipedia.org) holds scale-free small-world properties ((Zesch & Gurevych 2007); (Masucci et al. 2011)). Therefore we think that the knowledge structures represented in human mind might have some resemblance with information structures existing already currently in the Wikipedia online encyclopedia. We now report some experimental findings based on comparison we have carried out with word lists and conceptual networks generated by students, vocabulary ranking of British National Corpus, hyperlink network structure of the Wikipedia and exploration paths of students in the hyperlink network of the Wikipedia. A more extensive analysis and more detailed listings about experimental data and its comparisons are available in publication (Lahti, to appear). We propose a new educational framework for educational exploration based on conceptual networks generated and explored by students supplied with Wikipedia linkage.

Previous research

Finding the shortest route that visits each node in a network once then finally returning to start node again, known as a *travelling sales man problem*, has shown to be a NP-hard problem but interestingly human performance to solve travelling sales man problem has been shown to be close to optimal (Acuña & Parada 2010), thus motivating exploiting human-like intuitive heuristics for efficient exploration in a network.

Small-world networks are networks that have a small average distance (or diameter) between nodes d so that for N nodes in network each having z neighbors the average distance can be estimated with formula d = log N / log z (Newman 2000). Small-world networks have been considered as an interesting form of networks due to their flexible and efficient way to represent structure and growth of connectivity of various natural processes ((Watts & Strogatz 1998); (Kleinberg 2000); (Newman 2003)), and small-world networks have been identified in social networks (Uzzi et al. 2007), wikis (Mehler 2006) and the

Wikipedia online encyclopedia (Ingawale et al. 2009) as well as structurally and functionally in human brain networks (Wang et al. 2010).

Even when having very little knowledge of a given small-world network it has been shown that it is possible to route or navigate in it efficiently ((Kleinberg 2000); (Franceschetti & Meester 2006); (Sandberg 2008)). Liben-Nowell et al. (Liben-Nowell et al. 2005) showed that efficient decentralized search is enabled in social networks when relying on rank based friendship in which the probability of a person x having a person y as a friend is inversely proportional to the number of other persons being more closely positioned to x than y is.

Scale-free networks are networks whose nodes N have a probability of having k connections to other nodes that is proportional to $ck^{-lambda}$ with parameters c and lambda (Cohen & Havlin 2003). When parameter lambda in range $2 < lambda < 3$, average distance between nodes d in scale-free networks have been shown to be especially small following relation $d \sim \ln \ln N$ (Cohen & Havlin 2003). Wikipedia online encyclopedia holds scale-free small-world properties ((Zesch & Gurevych 2007); (Masucci et al. 2011)). Furthermore, scale-free properties have been possibly identified in human brain networks (Bullmore & Sporns 2009).

In a network when using a routing algorithm based on only local information, the number of nodes visited before reaching the target node is minimized when *probability of having a link* between two nodes decays with the square of their distance and only with this condition it is possible to reach the target in logarithmic number of steps (Franceschetti & Meester 2006). In networks among non-uniformly spaced nodes when rank(w) depicts ranking position of node w among all possible nodes linkable from node v, the linking from node v to node w has been suggested to have a probability $rank(w)^{-1}$ thus meaning probability decaying along ranking position (Easley & Kleinberg 2010).

Age of acquisition effect has been identified both in native language acquisition and secondary language acquisition meaning that words learned earlier in a person's life can be recognized and produced more quickly than words learned later in life and it has been suggested that mappings between orthographic, phonological and semantic representations of words form a network that support later reconfigurations for new associations but still favour connections learned early in language acquisition ((Izura & Ellis 2002); (Ellis & Lambon 2000)). *Word frequency effect* has been noted so that people respond more quickly to high-frequency words of a language than low-frequency words of a language in respect to for example lexical decision, reading aloud, semantic categorization and picture naming (Duyck et al. 2008).

As of February 2014, constantly growing *Wikipedia online encyclopedia* contains 4,5 million articles and possibly 2,1 billion hyperlinks based on relation suggested by (Zlatic et al. 2006) between the number of directed links L and articles N in the Wikipedia being approximately $L=N^{1,4}$. Each Wikipedia article can be considered as a concept defined by article title and hyperlinked articles can thus form *conceptual networks*. For each conceptual relationship it is possible to define a compact relation statement extracted from the sentence surrounding each hyperlink anchor in a Wikipedia article. *British National Corpus* (BNC) is a respected corpus containing 100 million words of samples of English language of which 90 percent is based on texts and 10 percent based on speech (Kilgarriff 1997). For reasonable comprehension it has been considered sufficient to understand 95 percent of general texts (Laufer 1989) which corresponds to a vocabulary of 3000–5000 or just 2000–3000 word families (Nation & Waring 1997).

It has been claimed that both personal and collective associative networks have a small-world structure and that collective associative networks either have a scale-free structure or do not have it ((Morais et al. 2013); (De Deyne & Storms 2008); (Steyvers & Tenenbaum 2005)). Olney et al. (Olney et al. 2012) found that the Wikipedia reflects the aspects of meaning that drive semantic associations concerning structure of language, organization of concepts/categories and linkage between them. According to a recall experiment with a semantic network model based on the Wikipedia (Thompson & Kello 2013) semantic memory processes can be usefully modeled as searches over scale-free networks and it was shown that inter-retrieval interval was progressively greater as minimum path length increased between nodes of semantic network to be recalled.

It has been suggested that learning effectiveness benefits from combined *distributed adoption and retrieval of knowledge* at the longest delay that still maintains correct recall ((Hunt & Beglar 2005) referring to (Landauer & Bjork 1978)). Based on previous research, Thalheimer (Thalheimer 2006) concludes that for recalling information successful learning experiments have had three or more repetitions and that longer spacing of repetition supports longer retention periods. About ten repetitions has been considered desirable to ensure learning a new word ((Nation & Wang 1999) referring to (Nation 1999)). With an assumption that weakening memory requires next encounter to be spaced at most by a

week, a suggestion was then formulated that a learner should read each week at least these same amounts of text ranging from 3226 words per week (on lower educational level) to 20000 words per week (on higher educational level) (Nation & Wang 1999).

Method

We propose a new educational framework to support learning with exploration in conceptual networks generated by students and exploration in corresponding hyperlink network of the Wikipedia as well as comparison of these knowledge structures. Based on previous research and our own research we especially suggest using cumulatively growing vocabularies and traversing the shortest paths of conceptual relationships connecting high-frequency words concerning new knowledge of learning topic and the learner's prior knowledge as well as exploiting principles of spaced learning for exposure and retention (see details in (Lahti, to appear)). We now report findings that we have gained with two different experimental setups with students, using Group 1 and Group 2, and there was no overlap of members between Group 1 and Group 2. The students had ages ranging from 15 to 19 years and had learning abilities that can be considered normal. They represented relatively diverse cultural backgrounds and school performance and some of them used in our experiment English language although a majority used Finnish language, but anyway we decided to report all our results in English.

In the first experimental setup we had Group 1 having 103 students to create high-frequency lists of words and links between words. We asked each student to freely associatively write a list of 20 most important concepts (only common nouns) concerning topic "life" (it was instructed that the concept "life" itself should not be mentioned in the list). Then we asked everyone to review his generated list and give each concept a ranking value representing "measure of importance" ranging from 1 to 20 (value 1 meaning the most important); we later translated each value to a descending range from 21 to 1 (value 21 meaning the most important). Then we asked each student to draw a concept map by adding in a free ordering all the concepts to a paper and connecting with a non-directional line the most important connections between these concepts according to her intuition (thus linking direction was not specified when defining relationships between a pair of concepts). Based on word lists, measures of importance and concept maps we tried to represent an approximated average conceptualization of knowledge of these students. Naturally, there are many alternative ways to define rankings for words and links.

The highest-ranking concepts in word lists based on occurrences were family (53), friend (49), work (41), death (40) and love/school (33), and the highest-ranking concepts in word lists based on sum of measures of importance are family (903), friend (821), love (525), work (445) and water (408). We contrasted word lists of Group 1 with the highest-ranking nouns of British National Corpus. From word lists of Group 1 we selected for further analysis only those words mentioned by at least four students thus ending up having a subset of 102 highest-ranking concepts, now called as "102 core concepts" (all of them belonging to word class of common nouns).[1]

From concept maps of Group 1 we decided to take into further analysis only those conceptual relationships existing between 102 core concepts and mentioned by at least two students in concept maps (we also included linkage to sister/brother since both sister and brother are represented by the same Wikipedia article Sibling) thus ending up having a subset of 145 conceptual relationships, now called as

[1] 102 core concepts (occurrences in word lists | sum of measures of importance), those indicated with an asterisk * belong to "hyperlink network of 55 concepts": air (9|121), animal* (29|285), baby (5|73), bed (4|44), biology* (5|44), birth* (23|321), book (10|99), bread (4|49), car (11|80), cat* (10|59), chair (4|10), child* (16|202), childhood (6|76), city (7|52), clock (9|98), cloth* (7|95), computer* (13|99), death* (40|363), disease* (6|28), dog* (15|118), dream_(sleeping) (4|53), eating (5|69), education* (14|172), elderness* (7|60), emotion* (6|86), environment (7|75), evolution (4|37), exam (4|30), experience* (6|66), family* (53|903), father* (7|105), flower (5|47), food* (31|396), forest (5|59), freetime* (7|91), friend* (49|821), fun (6|85), future (4|58), goal_(to_achieve) (4|64), god* (5|59), goodness (5|70), ground (6|74), growing (6|72), happiness* (11|179), hate (6|30), health* (14|225), heart* (6|80), hobby (15|188), holiday (7|91), home* (18|237), hospital (4|38), house* (15|147), human* (24|335), joy* (16|195), learning* (9|103), light* (7|67), living (8|105), love* (33|525), marriage (4|51), money (14|130), mother* (9|133), music* (8|91), nature* (21|303), nutriment* (4|61), organism* (4|41), oxygen* (4|79), paper (6|32), parent* (4|76), party (8|87), peace* (5|71), pen (5|33), people (4|41), pet* (7|64), philosophy (5|52), phone* (4|47), physical_training (4|54), plant* (12|136), pleasure (4|52), purpose (5|84), rain (4|34), religion* (8|62), sadness (4|32), school* (33|362), sea* (6|48), shoe (6|52), sister* (4|32), sorrow (14|104), sport (6|62), study (14|186), succeeding (5|64), summer (9|85), sun* (16|224), teacher* (4|34), television* (9|84), time (4|55), travel* (4|67), tree* (11|85), war* (5|27), water* (31|408), work* (41|445), world (7|106) and young_(person)* (5|30).

“145 core relationships”. Please note that the students did not specify linking direction for the relationships. Some of the highest-ranking conceptual relationships based on occurrences were family-friend (15), birth-death/family-love (13), friend-school (10), family-home/school-work (9) and animal-nature/friend-love (8). We contrasted conceptual relationships of Group 1 with a corresponding conceptual structure of hyperlink network of the Wikipedia (version as of 3 March 2008), now some of the 102 core concepts did not have directly corresponding article titles and then we tried to find the closest article title (air/Atmosphere_of_Earth, cloth/Clothing, elderness/Old_age, friend/Friendship, nutriment/Diet_(nutrition), physical_training/Physical_fitness, succeding/Management and young_(person)/Adolescence).

In the second experimental setup we had Group 2 having 49 students to explore a “hyperlink network of 55 concepts” that is such a subsection of hyperlink network of the Wikipedia (as of 3 March 2008) that had an overlap with 145 core relationships generated by students in the first experimental setup and still two additional constraints. These two constraints were the removal of concepts air and physical_training due to ambiguity and inclusion of only such concepts that were reachable in exploration paths (by traversing one or more intermediate hyperlinks) starting from concept human. Thus “hyperlink network of 55 concepts” consisted of altogether 212 hyperlinks connecting 55 nouns. In exploration each student could traverse each hyperlink at most once since it was not shown anymore if exploration later returned to the same start concept of hyperlink.

While students of Group 2 traversed 20 hyperlink steps in “hyperlink network of 55 concepts” on average 101,51 concepts (34,16 unique concepts) became shown to each student and each student selected 20 concepts (13,8 unique concepts). After the traversal each student could recall on average about 33,1 percent of unique shown concepts and on average 64,8 percent of unique selected concepts (see details in (Lahti, to appear)). We think that these results indicate that exploration in hyperlink network can support adoption of new knowledge thus enabling to develop new efficient learning methods and this motivates us to make the following comparative analysis about conceptual structures of students, Wikipedia linkage and exploration paths in it.

To compare distributions of frequencies and ranking values we used five statistical comparison tests: sign test of paired samples, bootstrap version of Kolgomorov-Smirnov two-sample test, Goodman-Kruskal gamma statistic, Spearman’s rank correlation coefficient rho and Kendall’s rank correlation coefficient tau. To facilitate identifying possible similarities between distributions we transformed – if needed – values into approximately same range of values thus forming scaled frequency distributions with empirically defined weighting parameters so that sign test of paired samples between a pair of distributions produces a p-value that is as high as possible and thus as an outcome the difference in medians between each three pairs of these scaled frequency distributions is as small as possible (see details in (Lahti, to appear)).

Experiment

Group 1 generated 621 unique nouns that had together 1777 occurrences, and among these 1777 occurrences 102 highest-ranking nouns had 1067 occurrences (60 percent of noun usage of Group 1) and 55 nouns of “hyperlink network of 55 concepts” had 772 occurrences (43 percent of noun usage of Group 1). Since getting 95 percent coverage of general texts suggested for reasonable comprehension can be possible with a vocabulary of just 2000–3000 word families ((Nation & Waring 1997) referring to (Laufer 1989)) we tried to estimate the coverage of our experimental vocabularies in lemmatized word list of British National Corpus (BNC) containing 6318 words occurring more than (or at least) 800 times in BNC (provided by Kilgarriff (Kilgarriff 1997), downloaded from (http://www.kilgarriff.co.uk/BNClists/lemma.num)). 102 highest-ranking nouns of BNC represented 5,8–6,0 percent among 2000–3000 highest-ranking concepts of any word class of BNC and 27–29 percent among 2000–3000 highest-ranking nouns of BNC, and respectively 55 highest-ranking nouns of BNC represented 4,0–4,1 percent and 18–20 percent. 55 nouns of “hyperlink network of 55 concepts” represented 1,0–1,1 percent among 2000–3000 highest-ranking concepts of any word class of BNC and 4,8–5,2 percent among 2000–3000 highest-ranking nouns of BNC.

We compared ranking difference (distance of ranking positions) based on sum of measures of importance (now on scale 1–21, greater value indicating more important) given by each student for the words she generated to form her word list and ranking based on occurrences in word lists generated by students. To enable comparison, the ranking values of sum of measures of importance given by students have now also been transformed into scale ranging from 1 to 102 (smaller value indicating more

important). Based on significance level of p<0,05, the null hypothesis was rejected in comparison tests of Kolgomorov-Smirnov, Goodman-Kruskal, Spearman and Kendall, and null hypothesis was not rejected in Sign test. Some of the greatest ranking differences for concepts having higher ranking based on sum of measures of importance given by each student than based on occurrences in word lists generated by students include (difference in parenthesis, suffix -s signifying shared positions): oxygen (+41,5s), parent (+40s), travel (+31s), goal_(to_achieve) (+27,5s) and purpose (+26s) and some of the greatest ranking differences for concepts having lower ranking based on sum of measures of importance given by each student than based on occurrences in word lists generated by students include: disease (-42s), hate (-40s), cat (-39,5s), paper (-37s) and city (-31s). Some of the smallest ranking differences include: family/friend/home (0), hobby/sun (0s), evolution (+0,5s) and clock/party (-0,5s).

We compared ranking difference (distance of ranking positions) based on concepts having higher ranking position for occurrences in British National Corpus than in word lists generated by students. Based on significance level of p<0,05, the null hypothesis was rejected in comparison tests of Kolgomorov-Smirnov, Spearman and Kendall, and null hypothesis was not rejected in Sign test and Goodman-Kruskal. Some of the greatest ranking differences for concepts having higher ranking based on occurrences in British National Corpus than based on occurrences in word lists generated by students include: time (+89,5s), people (+88,5s), parent (+62,5s), teacher (+61,5s) and bed (+57,5s). Some of the greatest ranking differences for concepts having lower ranking based on occurrences in British National Corpus than based on occurrences in word lists generated by students include: hobby (-73s), sorrow (-71s), joy (-63s), happiness (-58s) and human (-57). Some of the smallest ranking differences include: music (+0,5s), work (-1), philosophy/sadness/school (-1,5s) and health/sport (+2s).

Between 102 core concepts there were in concept maps drawn by students 145 relationships mentioned by at least two students (with 75 distinct concepts) and in the Wikipedia (as of 3 March 2008) 422 hyperlinks (with 93 distinct concepts). There were 69 shared distinct concepts in these 145 relationships of concept maps and 422 hyperlinks of the Wikipedia, and they had overlapping connectivity containing 44 relationships that is shown in Figure 1a.

We compared ranking difference (distance of ranking positions) based on 69 shared concepts having higher ranking position for occurrences as start/end nodes in hyperlinks of the Wikipedia than in relationships of concept maps drawn by students. Based on significance level of p<0,05, the null hypothesis was rejected in comparison tests of Kolgomorov-Smirnov, Spearman and Kendall, and null hypothesis was not rejected in Sign test and Goodman-Kruskal. Some of the greatest ranking differences for 69 shared concepts having higher ranking position for occurrences as start/end nodes in hyperlinks of the Wikipedia than in relationships of concept maps drawn by students include Oxygen (+51,5s), Religion (+49s), Biology (+45s), Adolescence/Emotion (+38,5s) and Plant (+35s). Some of the greatest ranking differences for 69 shared concepts having lower ranking position for occurrences as start/end nodes in hyperlinks of the Wikipedia than in relationships of concept maps drawn by students include Work (-61s), School (-47,5s), Joy (-44,5s), Friendship (-43,5s) and Home (-40,5s). Some of the smallest ranking differences include Happiness/Love/Sun (0s), Death (-1,5s), Child (+2s) and Disease (-2s).

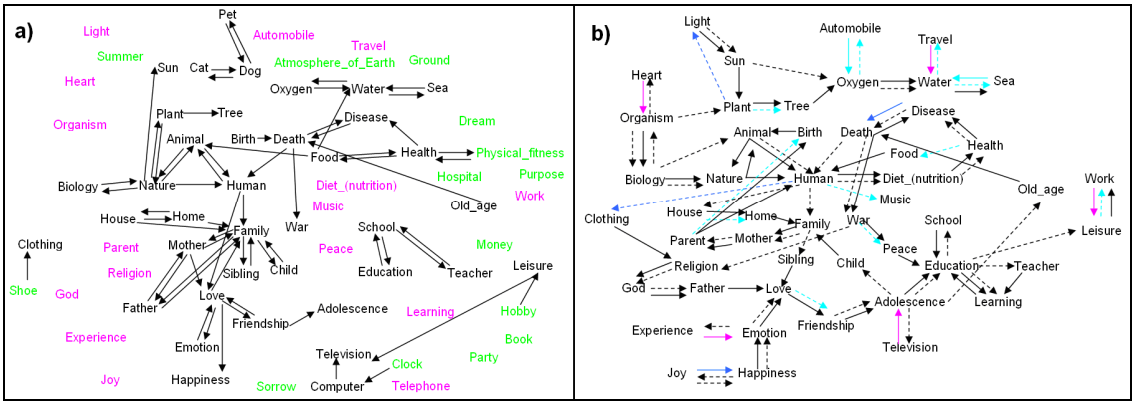

Figure 1. a) Overlapping connectivity between 69 shared concepts in concept maps drawn by students (n=103) and hyperlink network of the Wikipedia, containing 44 relationships (concepts with black or pink font belong to “hyperlink network of 55 concepts”). b) The most actively traversed departing links (solid

lines) and arriving links (dotted lines) in "hyperlink network of 55 concepts" (only these 49 of 55 concepts became visited in exploration).
Figure 1b shows the most actively traversed departing links (solid lines) and arriving links (dotted lines) in "hyperlink network of 55 concepts", in some cases more than one link shared the highest activity. Pink links are not in original "hyperlink network of 55 concepts" but are needed to roll back from dead ends in exploration. Turquoise links were inherently sole connecting arriving/departing link between these two concepts whereas blue links emerged as sole connecting links after roll back links had been excluded between these two concepts. Among all 55 concepts five concepts did not have any traversed arriving/departing linking, including Cat, Computer, Dog, Pet and Telephone, and concept Music had only traversed arriving link and not departing link.

We think that various forms of interactive and engaging learning activities can be developed based on the student's exploration in hyperlink network. To illustrate pedagogic potential of associative chaining of browsed concepts and relation statements (extracted from the sentence surrounding hyperlink anchor) in exploration paths we generated examples based on Figure 1b. An exploration path starting from concept Human and proceeding the most actively traversed departing hyperlinks in "hyperlink network of 55 concepts" generates the following learning path: Human->Diet_(nutrition)->Health->Disease->Death->War->Peace-> Education->Learning->Education (and then remaining in an eternal cycle Education->Learning-> Education->etc.). When chaining relation statements of each of these hyperlinks we gain an educational story shown in Figure 2a.

We think that even if having somewhat limited scope, already these examples show that suggested method of traversing exploration paths can offer to the student a relatively intuitive way to adopt step by step new pieces of knowledge in a simple process. Relying on exploration experiment with 49 students this exploration path can be considered to represent some kind of average association chain of students about gradually evolving thinking when starting from concept Human and finally reaching limits of this expansion when arriving to a repeating cycle. We believe that with sufficiently large and diverse collection of traversed exploration paths a student can achieve relatively extensive coverage of hyperlink network of concepts about desired learning topic. We think that this gained collection of exploration paths can offer interesting insight to the student's conceptualization and personal characteristics as well as to the semantic properties of language and consciousness.

Different perspectives can be achieved if exploration path proceeds a chain of arriving links instead of departing links. An exploration path starting from concept Human and proceeding the most actively traversed arriving hyperlinks in "hyperlink network of 55 concepts" generates two alternative learning paths since it appears that there are two most actively traversed arriving links arriving to concept Human that share the highest ranking and thus two different paths emerge proceeding to Death or Animal. One of these two paths is: Human<-Death<-Disease<-Health<-Diet_(nutrition)<-Human (and then again possibility to proceed to Death or Animal, i.e. leading to consecutive cycles that arrive back to Human or then leading to a path proceeding through concept Animal as explained next). The other one of two paths is: Human<-Animal<-Biology<-Organism<-Biology (and then remaining in an eternal cycle Biology<-Organism<-Biology <-etc.).

These just shown learning paths can be contrasted with a learning path generated based on the highest-ranking relationships in concept maps drawn by students (n=103) mentioned by at least two students and considering only those relationships that contain concepts belonging to 55 concepts of "hyperlink network of 55 concepts". When traversing relationships of concept maps (linking direction was not specified in relationships of concept maps) so that we start from concept "human" and proceed at each step to relationship that has the highest number of occurrences we get a learning path: human–family–friend–school–work–education–school (and then remaining in an eternal cycle school–work–education–school–etc.).

When comparing learning path generated based on relationships of concept maps with learning path generated based on "hyperlink network of 55 concepts" it seems that learning path based on relationships of concept maps focuses on social themes whereas learning path based on "hyperlink network of 55 concepts" focuses on survival themes. Anyway, interestingly learning paths based on relationships of concept maps and "hyperlink network of 55 concepts" with departing hyperlinks finally arrive to an eternal cycle having a shared theme concerning education. Further experiments with much bigger samples are needed to make more accurate estimates.

In respect to traversing exploration paths in networks shown in Figure 1b it could be also possible to select paths so that the highest-ranking concept based on various properties (for example the number of occurrences as start concept or end concept in hyperlinks as well as the number of occurrences in exploration paths) could be prioritized even when having distance longer than just one hyperlink.

Therefore each concept could be considered metaphorically to have some kind of own gravitational field and the sum of all these gravitational fields would then contribute to selecting at each step the next hyperlink to be traversed next in the hyperlink network.

We compared ranking difference (distance of ranking positions) based on 55 concepts of "hyperlink network of 55 concepts" in respect to occurrences as start/end concepts in hyperlinks of the Wikipedia and number of departures/arrivals from/to a concept in exploration paths of students (more than one departures/arrivals per concept can be counted for each student). These comparisons were based on significance level of $p<0,05$. When comparing number of departures from a concept and occurrences as start concept the null hypothesis was rejected in comparison tests of Goodman-Kruskal, Spearman and Kendall, and null hypothesis was not rejected in Sign test and Kolgomorov-Smirnov. When comparing number of arrivals to a concept and occurrences as end concept the null hypothesis was rejected in comparison tests of Goodman-Kruskal, Spearman and Kendall, and null hypothesis was not rejected in Sign test and Kolgomorov-Smirnov. When comparing number of departures from a concept and occurrences as end concept the null hypothesis was rejected in comparison tests of Kolgomorov-Smirnov, Goodman-Kruskal, Spearman and Kendall, and null hypothesis was not rejected in Sign test. When comparing number of arrivals to a concept and occurrences as start concept the null hypothesis was rejected in comparison tests of Spearman and Kendall, and null hypothesis was not rejected in Sign test, Kolgomorov-Smirnov and Goodman-Kruskal.

In "hyperlink network of 55 concepts" in respect to occurrences some of the greatest ranking start concepts for hyperlinks include Human (16), Education/Food (10) and Plant/Water (8), and some of the greatest ranking end concepts for hyperlinks include Human (11), Family (10) and Biology/Oxygen (9). In "hyperlink network of 55 concepts" in exploration paths some of the concepts having greatest number of departures include Human (90), Emotion (70) and Happiness/Love (50), and some of the concepts having greatest number of arrivals include Happiness (55), Emotion/Love (48) and Education (46).

In "hyperlink network of 55 concepts" in respect to occurrences in exploration paths some of the greatest ranking start concepts for traversed hyperlinks include Human (14), Education/Plant (8) and Water (7), and some of the greatest ranking end concepts for traversed hyperlinks include Human/Oxygen (9), Family (8) and Animal/Biology/Education/Love/Organism/Plant (6). In "hyperlink network of 55 concepts" in exploration paths some of the greatest ranking encountered concepts (at most one encounter per concept can be counted for each student) include Love (30), Emotion/Human (28), Experience/Happiness (26), Adolescence (25), Biology/Family (23). When examined separately for male and female students, some of the greatest ranking encountered concepts for males (n=18) include Human (12), Diet_(nutrition) (9) and Animal/Biology/Death/Disease/Experience/Love/Organism/Oxygen/Plant (8), and for females (n=31) include Emotion/Love (22), Adolescence (20), Happiness (19), Experience (18) and Family (17).

We compared ranking difference (distance of ranking positions) based on concepts having higher ranking position for occurrences in word lists of students than for encountered concepts in exploration in "hyperlink network of 55 concepts" (here an asterisk * indicates a concept that did not become encountered in exploration). Based on significance level of $p<0,05$, the null hypothesis was rejected in none of the comparison tests, and null hypothesis was not rejected in Goodman-Kruskal, Spearman and Kendall. Some of the greatest ranking differences for concepts having higher ranking position for occurrences in word lists of students than for encountered concepts in exploration include Food (+42s), Work (+38s), Dog* (+35,5s), Home (+34s) and Computer* (+32s). Some of the greatest ranking differences for concepts having lower ranking position for occurrences in word lists of students than for encountered concepts in exploration include Organism (-41s), Adolescence/Diet_(nutrition) (-39s), Biology/Emotion (-37,5s), Experience (-35,5s) and Parent (-34,5s). Some of the smallest ranking differences include Joy (-0,5s), Religion (+1s), Telephone* (+1,5s), Leisure/Sea (-3s) and God/Learning/Mother (-4s).

We compared ranking difference (distance of ranking positions) based on concepts having higher ranking position for sums of measures of importance given by each student than for encountered concepts in exploration in "hyperlink network of 55 concepts" (here an asterisk * indicates a concept that did not become encountered in exploration). Based on significance level of $p<0,05$, the null hypothesis was rejected in comparison tests of Goodman-Kruskal, Spearman and Kendall, and null hypothesis was not rejected in none of the comparison tests. Some of the greatest ranking differences for concepts having higher ranking position for sums of measures of importance given by each student than for encountered concepts in exploration include Food (+43,5s), Work (+37s), Home (+34s), Birth (+31s) and Dog* (+30s). Some of the greatest ranking differences for concepts having lower ranking position for sums of measures of importance given by each student than for encountered concepts in exploration include Adolescence (-

47), Biology/Disease (-41,5s), Organism (-39,5s), Experience (-35,5s) and Diet_(nutrition) (-30,5s). Some of the smallest ranking differences include Health/Light (+0,5s), Learning (-1s), Mother (+2s), Love/Peace (-2) and Joy (-2,5s).

Some of the highest-ranking traversed hyperlinks of the Wikipedia in exploration paths of students in "hyperlink network of 55 concepts" (number of traversals in parenthesis, here an asterisk * indicates a hyperlink that do not exist among core relationships of concept maps drawn by students) include Happiness->Emotion* (29), Emotion->Love (26), Joy->Happiness* (24), Disease->Death (24), Happiness->Joy* (21) and Human->Diet_(nutrition)*/Emotion->Experience* (19). Figure 2b shows in "hyperlink network of 55 concepts" all those links shared by both core relationships of concept maps drawn by students (blue lines) and traversed hyperlinks of the Wikipedia in exploration paths of students (red lines, dotted lines with arrow if traversed unidirectionally and solid lines if traversed bidirectionally). Greater width of line indicates higher position in ranking among those core relationships of concept maps and traversed hyperlinks that are shared by both listings, and the range of line widths is normalized for both listings to enable direct comparability.

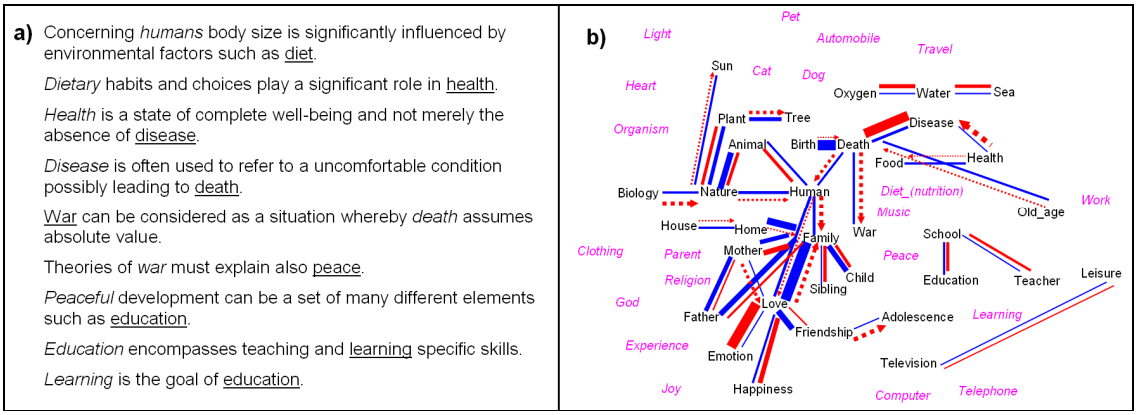

Figure 2. a) An educational story based on chaining relation statements of traversed hyperlinks (start concept of hyperlink indicated with italics and end concept of hyperlink with underlining). b) In "hyperlink network of 55 concepts" all those links shared by both core relationships of concept maps drawn by students and traversed hyperlinks of the Wikipedia in exploration paths of students.

Discussion and future work

We think that high-frequency word lists and high-frequency link lists enable to define a conceptual frame for the knowledge structures typically needed in education. We propose a new educational framework for educational exploration based on conceptual networks generated and explored by students supplied with Wikipedia linkage. Due to emergence of small-word properties and possibly scale-free properties in human brain and the Wikipedia linkage we think that these compact network structures can be promising for representing educational knowledge and processes. Our comparative analysis aims to show similarities and differences that we have identified when comparing knowledge structures generated and explored by students supplied with Wikipedia linkage. A more extensive analysis about our results is available in publication (Lahti, to appear). We hope that our findings can helps to define new kind of adaptive educational processes and identify requirements for setting the learning goals that can rely on exploiting large knowledge resources available in the Wikipedia an other open knowledge. We expect that every group of students will naturally generate somewhat different average high-frequency word lists and average high-frequency link lists. Especially we expect that along the learning process and maturing of student these lists can be seen evolving and possibly there are some shared trends of evolution and possibly these lists reach towards a conceptualization that can be considered to be somewhat a consensus of grown-ups about viewpoint on life. However we expect that in accordance with idea of life-long learning the evolution of these lists remain active through an individual's whole life enabling her always to excel herself further.

In our research we decided to emphasize analysis on teenaged students but we believe that our findings and modeling that we make with this age group can to some extent apply for students in other age groups as well. One of the reasons to emphasize teenaged students was that we expected that in our

experiments it was more easy to reliably convey the goals of our educationally motivated empirical tasks to relatively mature students than younger students (or younger children) and then to evaluate and model more reliably their learning processes.

When learning relies on exploration in hyperlink network we think that finding the most educationally rewarding path can be supported also with solutions identified for optimal stopping procedure (i.e. marriage problem, also concerning Odds algorithm) and related to this it has been found that brain regions identified to take part in evidence integration and reward representation encode threshold crossings which trigger decisions about committing to choice (Costa & Averbeck 2013). Therefore while deciding among all n outgoing hyperlinks which outgoing hyperlink to traverse next from current concept and if learner must select or reject each of alternative outgoing hyperlinks one by one, we suggest that optimal strategy can be to first directly reject about n/e of alternatives (here e denotes Napier's constant) and then select the next alternative that is better than all alternatives so far (or to select the last alternative) thus leading to that the probability of selecting the best alternative converges towards 1/e (≈ 0.3679) when n increases, as motivated by results of Bruss (Bruss 1984).

Future research needs to get better understanding about how students in a real educational setting traverse intuitively in hyperlink network of the Wikipedia. It is important to try to identify pedagogically fertile features of associative conceptual chaining in diverse exploration paths. Future research should carry out exploration experiment with students in such a hyperlink network that has a sufficient coverage and compactness and hopefully it could be also usefully contrasted to our experimental data about conceptual learning.

References

Acuña, D., & Parada, V. (2010). People efficiently explore the solution space of the computationally intractable traveling salesman problem to find near-optimal tours. Public Library of Science ONE (PLoS One), 5(7), e11685.

Bruss, F. (1984). A unified approach to a class of best choice problems with an unknown number of options. Annals of Probability, 12(3), 882-889.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 10(3), 186-198.

Cohen, R., & Havlin, S. (2003). Scale-free networks are ultrasmall. Physical Review Letters 90(5):058701.

Costa, V., & Averbeck, B. (2013). Frontal-parietal and limbic-striatal activity underlies information sampling in the best choice problem. Cerebral Cortex. Doi: 10.1093/cercor/bht286. First published online 18 October 2013.

De Deyne, S., & Storms, G. (2008). Word associations: network and semantic properties. Behavior Research Methods, 40, 213-231.

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. (2008). The frequency effect in second-language visual word recognition. Psychonomic Bulletin & Review, 15(4), 850-855.

Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press. http://www.cs.cornell.edu/home/kleinber/networks-book/

Ellis, A., & Lambon R. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. Journal of Experimental Psychology: Learning, Memory and Cognition, 26, 1103-1123.

Franceschetti, M., & Meester, R. (2006). Navigation in small-world networks: a scale-free continuum model. Journal of Applied Probability, 43(4), 1173-1180. Applied Probability Trust. http://fleece.ucsd.edu/~massimo/Journal/JAP-SmallWorld.pdf

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. Reading in a Foreign Language, 17(1), ISSN 1539-0578. http://nflrc.hawaii.edu/rfl/april2005/hunt/hunt.html

Ingawale, M., Dutta, A., Roy, R., & Seetharaman, P. (2009). The small worlds of Wikipedia: implications for growth, quality and sustainability of collaborative knowledge networks. Proc. Americas Conference on Information Systems (AMCIS 2009).

Izura, C., & Ellis, A. (2002). Age of acquisition effects in word recognition and production in first and second languages. Psicológica, 23, 245-281. http://www.uv.es/revispsi/articulos2.02/4.IZURA%26ELLIS.pdf

Kilgarriff, A. (1997). Putting frequencies in the dictionary. International Journal of Lexicography 10(2), 135-155. (A companion web site: BNC database and word frequency lists by Adam Kilgarriff. Lemmatised frequency list for 6318 words having more than (or at least) 800 occurrences in 100 million words of British National Corpus (BNC). Online available at http://www.kilgarriff.co.uk/bnc-

readme.html and http://www.kilgarriff.co.uk/BNClists/lemma.num)
Kleinberg, J. (2000). The small-world phenomenon: an algorithmic perspective. Proc. 32nd annual ACM symposium on theory of computing (STOC), 163–170. ACM Press. http://www.cs.cornell.edu/home/kleinber/swn.pdf
Lahti, L. (to appear). Computer-assisted learning based on cumulative vocabularies, conceptual networks and Wikipedia linkage. Doctoral thesis (submitted for evaluation in January 2014), Department of Computer Science and Engineering, Aalto University School of Science, Finland. spbvpaaj pwjpy xcbbl xgka xpyygbbj ölvvlcbbl vttbzwwtayl xctablpy.
Landauer, T., & Bjork, R. (1978). Optimum rehearsal patterns and name learning. In M. M. Grunberg, M., Morris, P., & Sykes, R. (eds.), Practical aspects of memory. Academic Press, London, UK.
Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In Lauren, C. & Nordman, M. (eds.), Special Language: From Humans Thinking to Thinking Machines. Multilingual Matters, Clevedon, UK.
Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. Proc. National Academy of Sciences (PNAS), 102(33), 11623-11628.
Masucci, A., Kalampokis, A., Eguíluz, V., & Hernández-García, E. (2011). Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. Public Library of Science ONE (PLoS ONE), 6(2), e17333.
Mehler, A. (2006). Text linkage in the wiki medium - a comparative study. Proc. Workshop on NEW TEXT - wikis and blogs and other dynamic text sources, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). http://acl.ldc.upenn.edu/eacl2006/ws12_newtext.pdf
Morais, A., Olsson, H., & Schooler, L. (2013). Mapping the structure of semantic memory. Cognitive Science, 37, 125-145.
Nation, P., & Wang, M. (1999). Graded readers and vocabulary. Reading in a Foreign Language, 12, 355-379. http://nflrc.hawaii.edu/rfl/PastIssues/rfl122nation.pdf
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.
Nation, I. (1999). Learning vocabulary in another language. E.L.I. occasional publication number 19, LALS, Victoria University of Wellington, New Zealand.
Newman, M. (2000). Models of the small world. Journal of Statistical Physics, 101(3/4), 819-841.
Newman, M. (2003). The structure and function of complex networks. Society for Industrial and Applied Mathematics (SIAM) Review, 45(2), 167-256. http://arxiv.org/pdf/cond-mat/0303516.pdf
Olney, A., Dale, R., & D'Mello, S. (2012). The world within Wikipedia: an ecology of mind. Information, 3, 229-255. doi:10.3390/info3020229
Sandberg, O. (2008). Neighbor selection and hitting probability in small-world graphs. The Annals of Applied Probability, 18(5), 1771-1793. http://www.cs.brown.edu/courses/csci2531/papers/AAP-Sandberg-NeighborSelect.pdf
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognitive Science, 29, 41-78.
Thalheimer, W. (2006). Spacing learning over time: what the research says. Will Thalheimer, Work-Learning Research Inc., Somerville, Massachusetts, USA (published March 2006, reformatted 2010). Online available at http://willthalheimer.typepad.com/files/spacing_learning_over_time_2006.pdf
Thompson, G., & Kello, C. (2013). Searching semantic memory as a scale-free network: evidence from category recall and a Wikipedia model of semantics. Proc. 35th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Austin, TX, USA.
Uzzi, B, Amaral, L., & Reed-Tsochas, F. (2007). Small-world networks and management science research: a review. European Management Review, 4, 77–91. EURAM Palgrave Macmillan Ltd.
Wang, J., Zuo, X., & He, Y. (2010). Graph-based network analysis of resting-state functional MRI. Frontiers in Systems Neuroscience, 4:16.
Watts, D., & Strogatz, S. (1998). Collective dynamics of "small world" networks. Nature, 393, 440–442.
Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. Proc. TextGraphs-2 Workshop at the 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics concerning Human Language Technologies (NAACL-HLT 2007).
Zlatic, V., Bozicevic, M., Stefancic, H., & Domazet, M. (2006). Wikipedias as complex networks. Physical Review E 74, 016115 (2006). http://cdsweb.cern.ch/record/931270/files/0602149.pdf?version=1