Learning to find interesting connections in Wikipedia

Marek Ciglan[°], Étienne Rivière[‡°], Kjetil Nørvåg[°]
[°] *Dept. of Computer and Information Science, NTNU, Trondheim, Norway*
[‡] *Computer Science Department, Université de Neuchâtel, Switzerland*
Email: {marek.ciglan,kjetil.norvag}@idi.ntnu.no, etienne.riviere@unine.ch

***Abstract*—To help users answer the question, what is the relation between (real world) entities or concepts, we might need to go well beyond the borders of traditional information retrieval systems. In this paper, we explore the possibility of exploiting the Wikipedia link graph as a knowledge base for finding interesting connections between two or more given concepts, described by Wikipedia articles. We use a modified Spreading Activation algorithm to identify connections between input concepts. The main challenge in our approach lies in assessing the strength of a relation defined by a link between articles. We propose two approaches for link weighting and evaluate their results with a user evaluation. Our results show a strong correlation between used weighting methods and user preferences; results indicate that the Wikipedia link graph can be used as valuable semantic resource.**

I. Introduction

Wikipedia is a free on-line encyclopaedia, created by a massive collaboration of volunteers from all over the world. It is currently the largest encyclopaedia in existence. Its scale and wealth of the information has been recognized by a number of scientists as a valuable source of the data for research in many domains of computer science. Wikipedia has been used for solving natural language processing tasks, for enriching information retrieval systems, and for ontology building. It is not only the sheer scale of Wikipedia that is of interest; it's also its structure. Each article is dedicated to a single topic and articles are densely linked among themselves. In addition, articles can be assigned to categories and Wikipedia categories form a hierarchy of its own. All of these properties have been exploited in research in recent years, using the textual content or the structural information – link graph and category graph.

In this work, we use Wikipedia's structure to answer the question, 'what is the connection between two or more concepts'. The motivation is to find out what is common and what connects given input concepts (in this paper, we use the term concept to refer to a subject described by a Wikipedia article). For example: *What is the connection between comedians Monthy Python and Peter Sellers? What is the connection between musicians Iggy Pop, Nick Cave and Dinosaur Jr.?* We aim to exploit information encoded in encyclopaedia hyperlinks to provide user with the answers.

We propose to use the modified Spreading Activation algorithm (SA) on the Wikipedia link graph to identify interesting connections between concepts. The rationale behind this decision is the following: as was shown by Zesch and Gurevych in [1], the Wikipedia link graph exhibits small-world and scale-free properties also common to semantic networks. The Spreading Activation algorithm was designed for searching semantic and association networks. A straightforward approach, in which we use the SA on the Wikipedia link graph with constant weights of the edges, show that we can use this approach to retrieve connection between concepts. However, results of initial experiments were very general, often obvious and of little interest to the users. To overcome this drawback, we propose two approaches for weighting links in the Wikipedia link graph and evaluate them using user judgments.

The main contributions of this paper are two approaches for weighting the strength of relations defined by links in Wikipedia. We evaluate them using an application for finding connections between Wikipedia topics; the results of the user evaluation show a strong correlation between user preferences and used weighting method. Results indicate that, although untyped, links in Wikipedia denote a semantic relationship, and that the node indegree in Wikipedia is an indicator of topic generality (first suggested in [2]).

The rest of the paper is organized as follows. After describing the related work (Section II), we define the challenge (Section III), describe our approach in Section IV, give a few implementation details in Section V and summarize results of the user evaluation in Section VI. We conclude the paper in Section VII where we also present directions for the future work.

II. Related work

In this section, we first summarize research works that utilizes the Wikipedia graph structure, and we briefly discuss the use of the SA in the context of Wikipedia research. Wikipedia graph structures, the link graph as well as the category graph, have been extensively used in research in recent years. Most notably, it was exploited for estimation of semantic relatedness (in [3], [4], [5] or in [6]), but it was used for other purposes as well – for extracting semantic relationships between categories [7], for building a thesaurus [8] and a taxonomy [9] or for improving ad-hoc information retrieval [10]. Extensive survey of Wikipedia uses in the research is provided by Medelyan et al. in [11].

Several authors have already applied the SA algorithm on the Wikipedia data; in [12] Syed et al. use the SA over Wikipedia to find generalized and common concepts related to a set of documents; in [13] Nastase et al. use the SA over Wikipedia to expand query terms for a summarization method and in [14] Waltinger and Mehler exploit the SA over the Wikipedia structure for context sensitive name entity recognition.

As was argued also by Kamps and Koolen in [10], links in Wikipedia define relations between articles. The type of the relation can be usually assessed quite easily by the human reader, by inspecting appropriate article texts. Knowing the relation types defined by links between articles would be very helpful in using the Wikipedia link structure as a full-scale semantic resource. Völkel et al. [1] have proposed and implemented a semantic Wiki system that enables the semantic typing of links. Although their approach is gaining popularity, for now, unfortunately, we do not have the luxury of having machine understandable descriptions of link types in Wikipedia. In this paper, we do not try to solve the problem of the Wikipedia link typing; we, however, propose two approaches for weighting links in Wikipedia. Our goal is to assess the strength of the relations defined by links between articles.

III. Problem statement

The goal of this work is to find interesting connections between two or more concepts (represented by articles) in Wikipedia, using its link graph and present them to the user. Under the term Wikipedia link graph, we understand an oriented graph created from Wikipedia articles, where nodes represent articles and edges represent links between articles.

We say that a concept c connects concepts a and b, if there is is a directed path from a to c and from b to c. We use the term connection to denote a triple (c, I, P) where I is a set of input concepts, c is a node that connects nodes in I and P is a set of paths that connects I and c. More formally, Let $W = (WV, WE)$ be the Wikipedia link graph, where WV is the set of all nodes (articles) and WE is the set of all edges in Wikipedia. Let $p(a, b)$ denotes a path from a to b. A connection is a triple (c, I, P) where $c \in WV; I \subseteq WV;$ *Pset of paths in W*$; \forall v \in I : \exists p(v, c) \in P \wedge (\forall p(a, b) \in P : a \in I \wedge b = c)$.

Let us mention that a direct path from a to b (where a and b are input nodes) is a special case of a connection where connecting node $c = b$ and path $p(b, c)$ is zero-length.

Such connections can be easily transformed into textual, human readable form – nodes can be represented by the titles of corresponding articles; edges can be described by the phrases from the article. For illustration, we provide an example of such a connection in textual form (from user evaluation described in Section VI); it represents a connection between concepts 'Michael Jackson' and 'Pink (singer)' where the connecting node is the concept 'Lisa Marie Presley':

Michael Jackson -> Lisa Marie Presley

spouse = Lisa Marie Presley (1994-1996)

Pink (singer) -> Lisa Marie Presley

... her good friend Lisa Marie Presley on the track Shine, on Presley's sophomore album "Now What"

The problem is that there are usually too many connections between input concepts; far too many for manual user inspection. For example, the average number of connections found after just two iterations of activation spreading among the pairs of concepts used in the evaluation was 5152. The question is, *how to identify connections that are likely to be interesting for the user and that should be presented to him*. Our approach is to use the Spreading Activation algorithm on the Wikipedia link graph to identify connecting nodes for a set of input concepts. Using this approach, the problem is reduced to the question, how to weight edges in the Wikipedia link graph, so that the weight represents the strength of their relation.

IV. Identifying connections in Wikipedia

In this section, we first present modifications to the standard SA technique that we use to identify connections for an input set of concepts. We then propose two approaches for weighting links of the Wikipedia link graph, which should help us to find interesting connections using the SA algorithm.

A. Modification of the standard Spreading Activation method

Spreading Activation is a method for associative retrieval from a graph data structure. Its basic form, as well as the most common variations, are extensively described in [15]. The nodes in graph structure represents objects and links denotes relations between the objects. Each node has an activation value, initially set to 0; each edge has a weight denoting the strength of the relation. The algorithm starts by setting activation of the input nodes to a predefined value. The processing consists of iterations, in which the activation is propagated from activated nodes to their neighbouring nodes. A node receives an activation in an iteration equal to the output of its neighbours weighted by the edge weights. Node output is typically computed using a threshold function, e.g., output is 0 if the activation level of the node is under threshold and 1 (or node activation value) if it is greater or equal than the threshold. Spreading of the activation continues until some termination condition is reached (e.g., number of iterations).

Mechanism of the SA utilize the breadth first expansion from activated nodes. This makes it a viable candidate for identification of connecting nodes. We can start the activation spreading from input nodes and identify the nodes with high activation, as nodes of interest. We have to solve

two problems – the identification of connecting nodes and the identification of paths from initial nodes.

In the standard SA algorithm, we can not distinguish whether a node received the activation from one or multiple initial nodes. We have modified the SA algorithm and store the node activation as a vector of float values (in the standard version the activation is modeled as a single value). The length of the activation vector is equal to the number of input concepts and the n-th value of the vector represents the amount of the received activation originated from the n-th input concept. The activation of the n-th input node is initiated as follows: all the values of the vector are equal to 0, expect n-th element that is initially set to 1.

Informally, the activation spread is computed individually for each input node. In addition to that, in each iteration, if the individual elements of node's input vector are lower than defined threshold t but the sum of all the elements is greater than t, we spread an output activation vector with a non-zero element, which is the element with highest value in the input activation vector.

Let $I_n = \{I_{n_1}, I_{n_2}, \ldots, I_{n_m}\}$ be the input activation vector of node n, where I_{n_i} is the amount of activation received from i-th initial node. The output function in the modified SA using the activation vector is:

$$O_n = \begin{cases} (thr(I_{n_1}, th), thr(I_{n_2}, th), \ldots, thr(I_{n_m}, th)) \\ \qquad iff\ \exists i : thr(I_{n_i}) \geq th \\ (m(I_{n_1}, I_n), \ldots, m(I_{n_m}, I_n)) \\ \qquad iff\ \nexists i : thr(I_{n_i}) \geq th \wedge \sum_i I_{n_i} \geq th \\ (0, 0, \ldots, 0) \\ \qquad iff\ \nexists i : thr(I_{n_i}) \geq t \wedge \sum_i I_{n_i} < th \end{cases}$$

$thr(x)$ and $m(x)$ are functions

$$thr(x,t) = \begin{cases} 1\ ;\ x \geq t \\ 0\ ;\ x < t \end{cases} \qquad m(x, I_n) = \begin{cases} 1\ ;\forall i : I_{n_i} \leq x \\ 0\ ;\exists i : I_{n_i} > x \end{cases}$$

This modification allows us to observe which sources the node received the activation from (non-zero values in the activation vector) and the amount of the activation from each source. The second problem is the identification of the paths from initial nodes. For this purpose, for each node we keep an array of predecessors. The n-th element of the predecessor array stores identifier of the predecessor node. The predecessor node is the one from which the highest activation from the n-th input node was first received. The process of the modified SA using activation vectors is depicted in Figure 1.

Those two modifications allows us to use the SA method to find connections, as defined in Section III. From the set of processed nodes, we select nodes that received the activation from all the initial nodes and order them according to their total activation. The total activation is equal to the sum of all the elements of the activation vector.

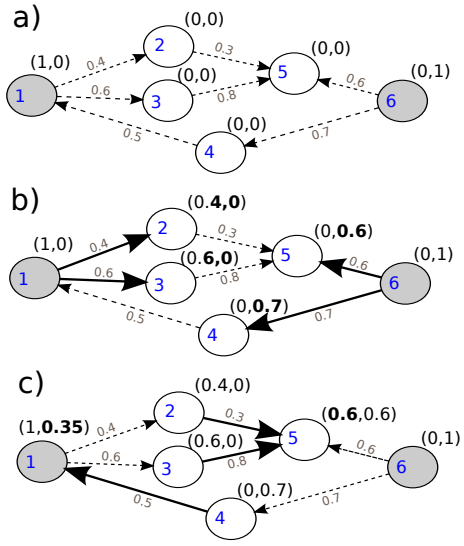

Figure 1. Modified SA with activation vectors (depicted above nodes); edge weights are above edges. a) initial state, only input nodes 1 and 6 are activated; b) first iteration, activation spreads from initial nodes; c) second iteration, nodes 5 and 1 are identified as connecting nodes Node 5 receives activation from multiple nodes (0.48 from node 3 and 0.12 from node 2, giving in total activation of 0.6 from input node 1). Node 5 keeps node 3 as a predecessor (for source 1), because it received the higher activation from node 3 than from node 2.

Connecting nodes of connections identified by this approach have high levels of activation, meaning they are, according to the information encoded in the graph structure, closely related to the set of input concepts. Having this mechanism in place, we have tried to use our modified SA algorithm on the Wikipedia link graph. In our first experiments, we have used activation decay and iteration length constraint (algorithm stops after defined number of iterations) and we have used the constant link weighting function (weights of all edges were the same). The results showed that we can use our approach to retrieve meaningful connection from the Wikipedia link graph. However, the results of this initial approach contained connecting nodes of a very general nature. For illustration we can take the example query from the introduction - What is the connection between Peter Sellers and Monty Python? Using the straightforward approach, top scoring connections contained connecting nodes such as BBS, England, London. While such general connections are perfectly correct, they are also quite obvious and often of little or none interest for the user. This leads us to the effort to try to suppress the obvious connections from top ranking results. Our approach was to modify the weights of the edges (i.e., weight of the relation between linked concepts) to achieve this effect.

B. *Edge-weighting approaches*

We have proposed two approaches for weighting edges in the Wikipedia link graph. The first one, named Indegree

Square Ratio, is based purely on graph properties. The second one is called CatTax, and our intention was to use semantics describing concepts, for the link weighting; it exploits a *is-a* taxonomy derived from Wikipedia category graph, developed by Ponzetto and Strube [9].

1) Indegree Square Ratio: The first presented edge-weighting approach is called Indegree Square Ratio (ISR). This approach uses purely graph properties and is thus applicable not only to Wikipedia but also to link graphs of other Wikies or, possibly, to other graph structures with similar properties. As was mentioned earlier, our intention was to minimize the appearance of very *general* connections, that are usually of a little interest to the user, among top ranking results. The challenge here is, how to identify whether an article is more general than other, based only on the graph properties of Wikipedia link graph. After inspection of the initial experiment results and the graph structure, we have observed that very general concepts tend to have a high indegree (number of incoming links). We thus followed the hypothesis that the node indegree is an indicator of its generality. A similar approach was also advocated by Gabrilovich and Markovitch in [2], where authors considered the article a to be more general than b iff $log_{10}(indegree(a)) - log_{10}(indegree(b)) > 1$.

We do not use the node indegree just to discriminate connecting nodes with high indegrees (e.g., indegree > threshold). This would corrupt the results of queries searching for the connections among general nodes. Instead, in our approach, we want to assign weights according to the generality of both nodes forming the edge. We want to assign high weights to the links that connect the nodes of the similar generality. Similarly, we want to assign high weights to the links where target node is less general than the starting node. The intuition behind this is that concepts referenced from more general concepts are likely to be important for the general concept; thus, the relationship should be strong. On the other hand, links oriented from less general concepts to more general concepts should be assigned with lower weights. The Indegree Square Ratio (ISR) weighting function for an edge $e(i,j)$ is:

$$ISR(e(i,j)) = \begin{cases} \frac{indegree(i)^2}{indegree(j)^2} & \textit{iff } \frac{indegree(i)^2}{indegree(j)^2} < 1 \\ 1 & \textit{iff } \frac{indegree(i)^2}{indegree(j)^2} \geq 1 \end{cases}$$

To illustrate the effect of this weighting, we present an example of connections identified by the SA constrained to only one iteration. The constrain of one iteration means that the only connections that can be identified for two input concepts a and b are: type 1: direct connections ($a \rightarrow b \vee b \rightarrow a$); type 2: three node connections of the form: $a \rightarrow c \wedge b \rightarrow c : c \in WV \wedge c \neq a \wedge c \neq b$.

The query is to find connections between rock music interprets *Iggy Pop* and *Joy Division*. There are in total 14 connections that can be identified with SA constrained to one iteration, one direct link and 13 connections of the type 2. Connecting nodes ordered according to their activation in the descendant order, using the constant edge weighting function are :

Ian Curtis; Punk rock; Post-punk; Rock music; David Bowie; RCA Records; Biographical film; Sex Pistols; Red Hot Chili Peppers; The Doors; Virgin Records; Control (2007 film); Jim Morrison

Very general connecting nodes, in this case music categories (e.g., Punk rock; Post-punk; Rock music) and record company (RCA Records) are ranked on top. Using the ISR weighting function, the connecting nodes ordered according to their activation are:

Control (2007 film); Ian Curtis; Jim Morrison; Biographical film; The Doors; Sex Pistols; Post-punk; Red Hot Chili Peppers; David Bowie; RCA Records; Virgin Records; Punk rock; Rock music

Using the ISR weighting function, very general concepts (music categories, recording companies) receive smaller amount of activation and are ranked at the end of the list. This is the behaviour we are trying to achieve – suppress the connections that are likely to be considered obvious.

2) CatTax: The ISR uses only the Wikipedia link graph to assess the strength of links. The second proposed edge-weighting approach is called CatTax. Our intention was to use the semantics describing the nodes to weight the strength of an edge connecting them. The more related the nodes are, the more weight should the edge connecting them have. We consider Wikipedia categories assigned to an article as a description of the node's content. We want to weight links according to the similarity of the nodes content. Thus, by measuring the relatedness between categories of two articles, we can obtain the weight of the link connecting them. As was shown in [5], Wikipedia category graph is a scale-free, small world graph and is similar in that respect to lexical semantic networks (e.g., WordNet). There exists a large number of semantic relatedness measures for lexical semantic networks and so, natural approach is to use one of them also for the Wikipedia category graph. Challenges for using semantics defined by categories to weight links between articles are the following:

- How to adapt semantic relatedness measures proposed for lexical semantic networks to the Wikipedia category graph.
- How to clean the Wikipedia category graph. As was shown by Chernov et al. in [16], links between Wikipedia categories have very different semantics, while some represent strong semantic relations, other are used only for navigation or unimportant purposes.

Both challenges have already been addressed by other authors. To adapt lexical semantic relatedness measures to Wikipedia category graph, we follow the approach proposed

by Zesch and Gurevych in [5]. We compute semantic relatedness of all pairs belonging to the cartesian product of categories assigned to two linked articles. We then choose the best value among all pairs to assess the weight of the link. The second problem is the cleaning of the Wikipedia category graph. We use the work presented in [9], where authors derived a taxonomy of $is-a$ relations from the raw Wikipedia category graph.

In order to measure semantic relatedness between a pair of categories we use the *simWP* measure introduced by Wu and Palmer in [17]. The measure exploits the lowest common subsumer – parent of both nodes with largest distance from the root node. The process of computing the weight of an edge $e(i,j)$ by CatTax measure is as follows:

1) Determine C_i and C_j – sets of categories assigned to nodes i and j.
2) Compute $SimWP(c_k, c_l)$ for all $c_k \in C_i \wedge c_l \in C_j$ over *is-a* taxonomy [9]
3) Select the highest value among all pairs (from step 2)

To illustrate the effect of this weighting, we present the results of the example query used in previous subsection. The connecting nodes are ordered according to their activation: *Post-punk; Ian Curtis; Sex Pistols; The Doors; David Bowie; Rock music; Control (2007 film); Biographical film; Punk rock; RCA Records; Red Hot Chili Peppers; Virgin Records*

Concepts representing artists and groups are grouped at the top of the list. The example query results indicates that the connecting nodes semantically related to the input concepts in the $is-a$ hierarchy are favoured.

C. Post-processing of results

Connections containing long paths are usually considered by users as quite obscure and vague. However, using our spreading activation approach, it is not unusual to receive a connection with a long path that has a significant activation value on the connecting node.

This happens when a node receives a large activation originated from a single source, which is sufficient to qualify it among top activated nodes and at the same time, it receives a marginal activations through a long paths from other initial nodes. The node is identified as a connecting node, because it receives activation from multiple sources and it is ranked high in the result list, as it receives large activation from one source. It this case, we receive a connection with high activation value which contains a long path.

To deal with the latter, we have introduced a post-processing step, in which we can modify the score of the identified connections. In our approach, we use a length filter to discriminate the connections containing one or more long paths. We use the ratio of the number of edges in the connection and maximal possible number of edges for the defined number of the SA iterations to normalise the activation of the connecting node.

V. Implementation

In this section, we provide few details on the implementation of the prototype service for connections retrieval from the Wikipedia link graph. In the data preparation step, we have parsed the Wikipedia dump (from 2008-07-27); we have used custom script, to extract link graph in form of adjacency list; we have also transformed links leading to the redirect pages, replacing the redirect pages with their targets.

We have implemented our modified spreading activation algorithm in Java programming language, the processing part is run as a service, it's functionality is exposed via XmlRpc interface. Earlier experiments with the data stored in a relational database on the disk leaded to long computation times even for small number of iterations. Thus, we keep the whole link graph in the memory for faster processing.

We provide two user interfaces for the system, one displays results in the textual form, the other presents the connection using graphical representations, as it is the natural way of presenting such graph-related data to users. We have implemented specialised graph layout for our application, in which a strong relation between two concepts (according to their activation vectors) result in these two concepts being represented close-by together, whereas two unrelated concepts lie at a distance from one another. This provide an intuitive and compact representation of the top resulting connections.

VI. Evaluation

The goal of this work is to identify connections between input concepts that have the potential to be interesting for the user and present connections to the user. It is in fact a recommendation system and, to our best knowledge, there are no standard test sets that we could exploit to evaluate the quality of the results. As for other recommendation systems, the quality of the results should be judged by the users. Thus, we have performed an user evaluation to assess the quality of proposed approaches. We have set tree main requirements for the user evaluation: evaluation should be short to complete (under 10 minutes), so that the users do not get de-motivated; input concepts should be of common knowledge, so the users are able to judge the relevance of presented connections and finally, evaluation should provide direct comparison between proposed weighting approaches and a random weighting.

We have prepared a user evaluation, in which we tried to balance the simplicity of interaction (so that users understand the process and are able to complete evaluation in a reasonable time) with the need to compare complex thing as the connections often are. Users were comparing three weighting approaches for connections retrieval from Wikipedia. Compared weighting methods were the two proposed ones (ISR and CatTax) and a Random Weighting

(RW) method. Results of all three approaches were post-processed as described above. The goal was to find out whether results of proposed approaches were more attractive for the users than connections retrieved by randomly assessing the weights of the links. Second goal was to compare proposed approaches (ISR and CatTax).

In the user evaluation, we use connections between pop/rock musicians, chosen as top search terms in rock&pop category from 2004 to August 2009, according to 'Google insights for search'[1]. We choose this approach to ensure that users are (at least partially) familiar with the input concepts. From the top search terms, we choose the terms representing names of interprets or groups. We have constructed the input pairs by the order they appeared in the ranking. E.g., the top four artists in our list were Michael Jackson, Pink (singer), The Beatles, Vanessa Hudgens; thus, the pairs for connection search were (Michael Jackson, Pink (singer)) and (The Beatles, Vanessa Hudgens). We have used first sixteen interprets from the list, forming eight input pairs that were used in the evaluation.

To ease the comparison, the results of the three compared weighting approaches were presented on a single screen, appearing in a random order. For each weighting method, top five connections were displayed in the evaluation application. Users were informed that results of distinct methods will be presented in a random order. The task of the users was to rate result sets produced by compared approaches. Rating of individual results lead to long time required for completing the user evaluation. This is why we chosed to rate result sets rather than individual results. Users were instructed to select result set containing the most obvious and/or most meaningless connections. After selecting the most obvious set, users were asked to select second most obvious from the two remaining option.

The reason why we instructed users to select the most obvious results rather than the most interesting is that interestingness is quite subjective; information that is interesting for one person might not be interesting for the other. Our assumption was that users can agree more on what is obvious rather than on what is interesting. In the context of music artist, we can consider connections to be obvious, if the two artists are connected through a concept that is common to large number of music artists. E.g., if the presented connection relates artists A and B through the concept 'guitar', meaning A plays 'guitar' and B plays 'guitar' as well, we might consider this relation to be quite obvious, as there are a lot of musicians that can play guitar. Following this approach, the weighting method that produces the results that are the least obvious and are not meaningless at the same time, have the best potential to provide user with results that would be interesting for him or her.

To keep the time to complete the evaluation short, we have used eight input pairs, displaying only top five connections produced by compared approaches. Because of the methodology for choosing the input pairs, used concepts sometimes had not much in common (e.g., in search for connections between artists active in different decades). In those cases, the top five connections were often quite similar, making the task of identifying obvious connections more difficult. Nevertheless, a strong pattern emerged from the user voting. We had fourteen people participating in the evaluation. The rest of this section summarizes the results.

Table VI shows the aggregated results, indicating how often were the result sets of different weighting approaches marked as the most obvious, second most obvious and the least obvious in the whole evaluation. For results interpretation, it is important to note that if the results were ranked randomly, all the values in Table VI would equal 0.33. The results shows that random weighting (RW) was identified as the most obvious in most of the cases (70%), while ISR was mostly voted as the least obvious option (61%).

Table I
AGGREGATED RESULTS OF USER EVALUATION

	RW	CatTax	ISR
The most obvious	0.70	0.20	0.11
2nd most obvious	0.23	0.48	0.29
The least obvious	0.07	0.32	0.61

We have also analysed the results for each input pair individually. For six input pairs (out of eight), the average ranking was (ordered from the most obvious option): RW, CatTax, ISR. In the two remaining cases the average user ranking was: CatTax, RW, ISR.

We take this average user ranking for each input pair and compared it with the ranking of users. The goal was to find out how close to the average ranking are the rankings of best performing users and the worst performing users. The best performing user is the one with the voting closest to the average case, the worst user is the one whose voting differs the most from the average case. The results are presented in Table VI, showing number of times the voting of the user did not match the average voting. The cases presented in the table are: two best fitting users, two worst fitting users, median difference (Med) and average difference (Avg).

Table II
USERS VOTING COMPARED TO AVERAGE VOTING

	Best user	2nd best	Worst user	2nd worst	Med	Avg
Most obvious	1	2	5	2	2	2.36
2nd most obvious	2	3	6	5	4	3.64
Least obvious	1	1	4	5	2	2.64

In case of a random picking, the difference from average case would be 5.33. The results of the worst performing user show numbers that we would expect from a random voting. However, the second worst user has already a good

[1]http://www.google.com/insights/search

correlation with average voting for the most obvious cases. One of possible explanations for the worst result might be that one of the evaluators has accepted to participate in the evaluation for the sake of being polite, but did not care enough to do a careful evaluation; instead, clicked his way through the evaluation application as fast as possible.

Our interpretation of the user evaluation results is the following: a strong pattern in user preferences emerged; from tree options obtained by different weighting approaches and presented to the users in the random order, users have in 70% of cases rated the connection set produced by RW as the most obvious. ISR was selected by the users as the best (the least obvious) in 61% of cases. CatTax was usually selected as the second most obvious method, but for two input pairs, in which cases it's results scored as the worst (most obvious/meaningless). Analysis of the user voting showed good correlation among users, with only one user whose voting was similar to a random selection.

VII. Conclusion and future work

The aim of this work was to help users find the answers to the question, what are the connections between given (real world) concepts. Our approach is limited to scope of Wikipedia; we have proposed method for identifying connections between a pair (or a set) of concepts, described by Wikipedia article. To achieve this goal and exploit the knowledge encoded in the Wikipedia link graph, we have used the modified Spreading Activation algorithm. The main challenge of our approach is the weighting of a strength of a relation defined by a link between Wikipedia's articles. We have proposed two approaches for link weighting, CatTax and ISR, and we have compared them together with a Random Weighting method in the user evaluation.

Even though the user evaluation was limited in its scope (we were comparing only connections between concepts from the domain of music artists) and in the number of participants, results are promising. Results showed strong correlation of the user voting and used methods. The ISR method has been identified as the best method out of the three compared, while Random Weighting was almost in all the cases voted as the most obvious one. As ISR uses only graph properties to weight edges, we plan to investigate the possibility of using it also for other Wikis and other graph structures with similar properties in future work.

References

[1] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, 2006.

[2] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *J. Artif. Int. Res.*, vol. 34, no. 1, 2009.

[3] D. Milne and I. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links." in *In Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*, 2008.

[4] D. Turdakov and P. Velikhov, "Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation," in *Proceedings of Colloquium on Databases and Information Systems (SYRCoDIS), 2008*, 2008.

[5] T. Zesch and I. Gurevych, "Analysis of the wikipedia category graph for nlp applications," in *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[6] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa, "Wikiwalk: Random walks on wikipedia for semantic relatedness," in *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, 2009.

[7] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantics relationships between wikipedia categories," in *Proceedings of the First Workshop on Semantic Wikis – From Wiki To Semantics*, 2006.

[8] M. Ito, K. Nakayama, T. Hara, and S. Nishio, "Association thesaurus construction methods based on link co-occurrence analysis for wikipedia," in *Proceeding of CIKM '08*, 2008.

[9] S. P. Ponzetto and M. Strube, "Wikitaxonomy: A large scale knowledge resource," in *Proceeding of the 18th European Conference on Artificial Intelligence, ECAI*, 2008.

[10] J. Kamps and M. Koolen, "The importance of link evidence in wikipedia," in *30th European Conference on IR Research, ECIR*, 2008.

[11] O. Medelyan, D. Milne, C. Legg, and I. H. Witten, "Mining meaning from wikipedia," *Int. J. Hum.-Comput. Stud.*, vol. 67, no. 9, 2009.

[12] Z. Syed, T. Finin, and A. Joshi, "Wikipedia as an Ontology for Describing Documents," in *Proceedings of the Second International Conference on Weblogs and Social Media*, 2008.

[13] V. Nastase, K. Filippova, and S. P. Ponzetto, "Generating update summaries with spreading activation," in *Proceedings of the Text Analysis Conference*, 2008.

[14] U. Waltinger and A. Mehler, "Who is it? context sensitive named entity and instance recognition by means of wikipedia," in *Proceedings of WI-IAT '08*, 2008.

[15] F. Crestani, "Application of spreading activation techniques in information retrieval," *Artif. Intell. Rev.*, vol. 11, no. 6, pp. 453–482, 1997.

[16] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantic relationships between wikipedia categories," in *In 1st International Workshop: SemWiki2006 - From Wiki to Semantics*, 2006.

[17] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994.