Heuristics for Semantic Path Search in Wikipedia

Valentina Franzoni[1], Marco Mencacci[1], Paolo Mengoni[1], and Alfredo Milani[1,2]

[1] Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong
{valentina.franzoni,milani}@dmi.unipg.it
{marco.mencacci,paolo.mengoni}@studenti.unipg.it

Abstract. In this paper an approach based on Heuristic Semantic Walk (HSW) is presented, where semantic proximity measures among concepts are used as heuristics in order to guide the concept chain search in the collaborative network of Wikipedia, encoding problem-specific knowledge in a problem-independent way. Collaborative information and multimedia repositories over the Web represent a domain of increasing relevance, since users cooperatively add to the objects tags, label, comments and hyperlinks, which reflect their semantic relationships, with or without an underlying structure. As in the case of the so called Big Data, methods for path finding in collaborative web repositories require solving major issues such as large dimensions, high connectivity degree and dynamical evolution of online networks, which make the classical approach ineffective. Experiments held on a range of different semantic measures show that HSW lead to better results than state of the art search methods, and points out the relevant features of suitable proximity measures for the Wikipedia concept network. The extracted semantic paths have many relevant applications such as query expansion, synthesis of explanatory arguments, and simulation of user navigation.

Keywords: heuristics search, semantic networks, collaborative networks, semantic similarity measures, random walk, information retrieval.

1 Introduction

In the era of Big Data, searching, browsing and collaboratively building online semantic networks are the typical tasks that absorb a lot of users activity e.g., in social networks, collaborative encyclopaedias, media sharing repositories et cetera. Collaboratively updating/adding a reference to a Wikipedia entry, or labelling a multimedia object in a repository in order to make it clearer, are examples of those *collaborative* actions. The action is called collaborative since the user, in order to share, inform or facilitate other users, purposely adds the content/relationship to the network.

The connections of a semantic collaborative network can provide important information about the semantic relationship among the network entities i.e., the nodes. In this work, we focus on the problem of finding a chain between two entries of the Wikipedia online encyclopedia i.e., a path of articles between a source and a target, by

B. Murgante et al. (Eds.): ICCSA 2014, Part VI, LNCS 8584, pp. 327–340, 2014.

following hyperlinks among articles. This problem is of great importance, since it can provide useful information concerning the subject entities, such as relationships and explanations. In particular, considering the notion as a *context* consisting of two concepts with corresponding Wikipedia articles, the intermediate nodes in the Wiki path or *semantic chain*, as well as the *surrounding nodes*, can represent *hidden concepts* or *underlying concepts*, relevant for a number of applications, such as natural language understanding, query expansion and automatic explanation.

Let consider for instance two Wikipedia entries like *Mars* and *Scientist* and let the path `Mars->planet->science->scientist`. The intermediate concepts in the chain can be used to generate an explanation of the relationship between *Mars* and *Scientist*, by focusing on meanings that are consistent with the underlying context, which will most probably link the concept of *Mars* as a *planet in the Solar System* rather than to *the ancient Greek god Mars*. The problem of the semantic chain search can be reduced to a problem of search in a graph. To establish which concepts a pair of terms implies in a dialogue in concept explanation, we can consider the path between them, where the starting and ending nodes in the path form the context.

The basic idea of our approach is to apply the *Heuristic Semantic Walk (HSW)* [17][18] framework, where a proximity measure *m*, defined between pair of concepts, and derived from the statistical results of a query [1][2][4] to a search engine *S*, is used as *heuristic*, and is applied to guide a path search the Wikipedia concept network. The *HSW* approach is a general framework which can be instantiated with different heuristics h_m based on a different proximity measure *m* (e.g. *confidence*, *Pointwise Mutual Information*, *Normalized Google Distance*) [2][3][4], used by different informed search algorithms. In general, the proximity measure *m* between two term t_1 and t_2 is computed by submitting simple queries to a search engine *S*, and using the statistics about frequency and co-occurrence of terms in the indexed resources. In other words, the measure $m(t_1, t_2)$ reflects the collective knowledge embedded in the search engine *S*, with respect to answer the question about how far away are concepts t_1 and t_2. An informed search algorithm *A* can then calculate the value of the heuristics for each candidate successors, and decide the direction where to expand the search.

The experiments, held on Wikipedia on a range of different semantic proximity measures, show that the proposed approach outperforms classical uninformed search methods. In particular, HSW with heuristic randomized search returns the path that connects two concept nodes in much faster times than an uninformed blind random search; moreover HSW returns a higher quality path, in a semantic point of view, than an uninformed blind search. This latter result is particularly important when the HSW is used for semantic applications e.g., in *query expansion*, where the nodes of the path are used as candidates for the query expansion.

This paper is organised as follows. In the second section, the main features of the proposed heuristic walk approach are described, and semantic walk strategies are considered in section 3, where the experimented proximity measures are also exposed. Conclusions are drawn and future directions of the research are finally discussed.

2 The Heuristic Semantic Walk Model

In the Heuristic Semantic Walk model, we consider is to browse a semantic network in order to connect a pair of concepts, formulated as the problem of searching paths between two nodes over an oriented graph.

Definition: a *semantic network graph* $\Sigma = (V, E)$, is defined by a pair where V is a set of vertices/concepts (e.g., the entries in Wikipedia), and $E \subseteq (V \times V)$ is a set of oriented edges, representing the links between concepts in the network (e.g., the anchor links in the text of a Wikipedia article toward a referenced article).

Definition: the semantic path finding problem or `Path(s,g)`, given a semantic network $\Sigma=(V, E)$ and two nodes $s, g \in V$, consists in finding if a sequence of vertices $(v_0, v_1,..., v_n)$ exists, such that $v_0=s$, $v_n=g$ and for each $i \in [0, n-1]$ *the edge* $(v_i, v_{i+1}) \in E$. Similarly.

Shortest Path search problem `SPath(s,g)` can be defined straightforwardly.

2.1 Shortest Path and Plan Quality

In the following of this paper, we will consider the *Semantic Path Extraction* problem as broadly equivalent to the *Shortest Path* extraction on Wikipedia, although it is not. Intuitively Wikipedia is a *network of concept definitions and explanations*, then the *shortest the path* between to concepts, should be also the more "meaningful" i.e., the shortest the path the more *direct* is the concepts relationships. In the real case Wikipedia linked entries can also contain *user introduced noise, personal* and *structural biases.* Although policies, guidelines and form of controls are in place, the users are completely free to arbitrarily modify a Wikipedia article, thus introducing unwanted *errors,* placing links on irrelevant concepts or on common terms thus influencing the *semantic quality* of a possible path. This and other problems, such structural biases and hub terms will be further discussed in this paper.

2.2 Semantic Proximity Measures as Search Heuristics

It is well known how path search in state space, could greatly benefit from an informed search strategy, but unfortunately there are not inherent properties of the problem at hand, which can be used to define heuristics using classical technique like problem constraints relaxation. In fact the Wikipedia network, or other collaborative concept networks, cannot straightforwardly be seen as a state space. In the case of a semantic network, the relaxation technique cannot be applied to define heuristics, since the node is not a state generated by an action. On the other hand, it can be observed that links among concepts are added by the collective collaborative effort of the users, with the purpose of providing further explanations and insight knowledge e.g., the Wikipedia linked articles, or providing explanations about the content e.g., labels and tags in a multimedia repository. In term of state space problem, we can say that it is possible move from a state $\underline{A}$ (or Wiki article) to another state B if a

is_related_to action is applicable from state *A* to state *B* i.e., if the community of users have decided it. The basic idea is to use a measure of *relatedness* and a heuristic. If we broadly assume that *relatedness* is monotonically non-decreasing, and we look at it transitively, then we can support the intuition that there is a high chance that following a path with higher relatedness to the goal will likely lead to the target goal. As candidate for such heuristics, we focus on *proximity measure* in the literature which can be calculated from statistical data extracted from collaborative collective sources of information, such as any search engine, that can be both a generalized one (e.g., Google, Bing) or an embedded one in specialized media repositories (e.g. YouTube, Flickr) or social networks (e.g. Facebook, Twitter). This approach will guarantee that the source of data reflects the dynamical judgment of the community of the engine/repository, and at the same time, it can be easily calculated by querying the engine. Semantic proximity measures have been widely used for semantic extraction and automatic clustering of terms [12] but, as far as we know, in the HSW model for the first time they were used as heuristics for semantic search.

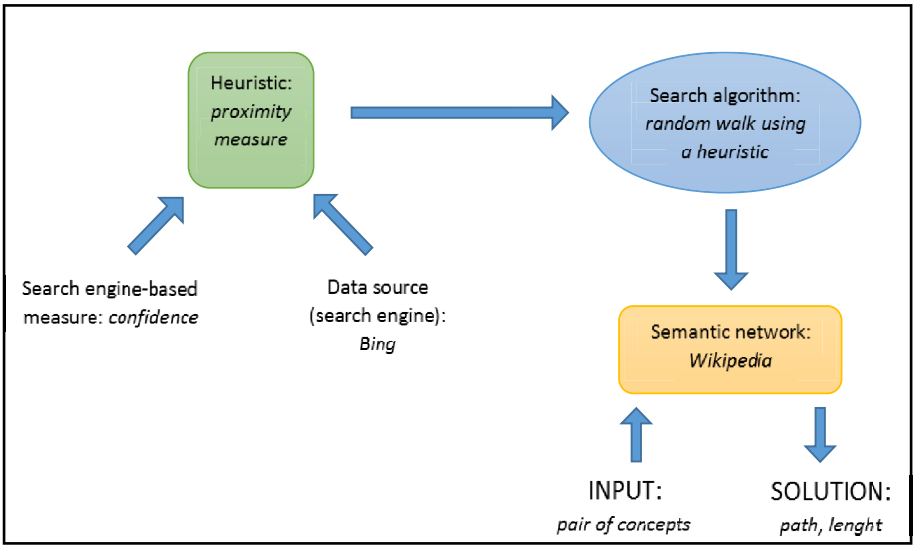


Fig. 1. Semantic Heuristic Walk General Scheme

2.3 Heuristic Semantic Walk (HSW): The General Scheme

The goal of a HSW is to return the path between a pair of terms *(s, g)* following the semantic links (e.g., the anchor links from the text of a starting Wikipedia article *s* to a goal article *g*), and using a semantic proximity measure to drive the search towards the best successor candidate (c_i) toward the goal node (*g*). A general scheme can be defined in order to characterize the class of HSW solutions to the semantic `SPath` problem.

Definition. A *Heuristic Semantic Walk instance*, or HSW, is an algorithm for `SPath` problems characterized by a *semantic collaborative network* graph $\Sigma=(V,E)$, a search engine S, which can return statistics about the occurrences of elements in $\wp(V)$, a proximity measure m defined in terms of those statistics, and a *search strategy* or *walk strategy A* using heuristics derived from m.

Note that $\wp(V)$ indicates the *part-of* set i.e., the set of all subsets of V; in other words, the engine/repository S should be able to return statistics for co/occurrences of objects. Most engines/repositories are in fact able to return statistics for queries like $Q=(a\wedge b\vee\neg c)$. Since the actual returned objects are not relevant, for our purposes the engine S can be formally defined.

Definition: a search engine/repository S is a function $S: \wp(V)\rightarrow\mathbf{N}$, where $S(Q)$ returns the number of occurrences of Q according to the engine/repository i.e., the statistics will reflect the frequency contextual use of terms in Q, according to the community related to the domain that feeds objects to S.

Given the previous general scheme, we will use the following notation, where $HSW(S,m,A)$ will denote an instance of HSW for a semantic network Σ where search strategy A uses heuristics h_m computed from source S, and $HSW(A)$ denotes a non-informed algorithm with a walk strategy A. For example, *HSW(Bing, NGD, A*)* for Wikipedia denotes an A^* search algorithm operating on Wikipedia, which uses as a heuristic the Normalized Google Distance computed using Bing statistics (see Fig.1).

3 Semantic Walk Strategies

The size and morphological features of online collaborative semantic networks pose important challenges and factors to be considered when choosing a walk strategy for a HSW scheme. In particular, the *high degree of nodes* i.e., the number of outgoing edges from semantic objects; the *high connectivity degree of the graph* i.e., the high number of *alternative paths* existing between a pair of nodes; and the nearly *unlimited size* of the network, point out drawbacks such as memory requirements, branching factors, endless loops and deadlock on optimality plateau, which can heavily affect the performance of the classical "best" graph search algorithms.

3.1 Non Informed Walks

Since the branching factor in Wikipedia can be very high, the BFS approaches are deeply penalized even in the case in which the target terms are only moderately distant from the source, by the exponential growth of the memory requirements. On the other hand, Depth First (DF) approaches cannot avoid to fall into loops, when re-encountering a node previously visited on a discarded branch, founding either loops or eventually very suboptimal paths [7][9][13] Iterative Deepening Search (IDS), which has low memory requirements and guarantees completeness, is extremely inefficient with high branching factors, and it is further penalized from the high time cost of the repeated online access to the candidate nodes. We can conclude that non informed

walk strategies cannot guarantee an efficient `SPath` search in the online collaborative semantic network scenario; on the other hand, complete algorithms based on IDS or BFS [5] still represent an important reference point, allowing to find benchmark optimal values for comparison purposes.

3.2 Heuristic Walks

Heuristic Walks are in general informed search strategies, which make use of a heuristic to estimate a score of each candidate node, and to sort them for the expansion. The evaluation is performed in terms of closeness/relatedness to the goal, where relatedness is computed by measuring the semantic proximity of the current node to the goal node. In the area of heuristic algorithms, we certainly have to consider A^* and algorithms derived from it, where its optimality properties would make it an ideal candidate for semantic heuristic walk strategies. The core of the A^* class of algorithms is the selection function $f(n)=c(n)+h(n)$ where $c(n)$ represents the cost to reach n from the source node (the path distance, in our case), while $h(n)$ is the heuristic estimate of the distance of n from the goal.

In A^*, a node n_e in the fringe F of an unexpanded node is selected for expansion if $n_e=min(n \in F, f(n))$ i.e., if the value of evaluation function is minimal for n_e. Unfortunately, as preliminary experiments have shown, when the target is not close to the source the A^* fringe F will grow too large with real online semantic networks, making the approach unpractical.

A *Greedy best-first* (GBF) approach has then been considered, since it avoids maintaining different nodes' path histories. In typical GBF, the evaluation $f(n)=h(n)$ does not consider previous costs, then only the nodes of the fringe that are the closest to the goal are chosen.

The problem of a growing large fringe have led to a particular implementation, *Local GBF* (LGBF), where no fringe memory is maintained except for the current path, and only the current node descendants are considered in the fringe. LGBF implements a local search with loop avoidance control.

3.3 Heuristic Weighted Random Walk (WRW)

In order to eliminate the problems with LGBF, we have decided to add a degree of randomization to the original LGBF. The basic idea of the semantic heuristic *Weighted Random Walk (WRW)* is to choose among a set of successors for the current node, making a random tournament among the candidate successors, weighted by the probability distribution induced on the candidates by the values of the proximity measure between the candidate concepts and the target concept. No memory of unexpanded nodes is maintained, except for the current solution path. In WRW, the proximity measure values need to be normalized to [0,1] in order to build the probability distribution. The probability of the candidate c_i in the random tournament is defined as

$$P(c_i) = \frac{m\,(c_i, g)}{\sum_j \; m\,(c_j, g)} \qquad \forall j \in Succ(curr) \tag{1}$$

where g is the target node, c_j are the successors of the current node *curr*, and $m(c_j,g)$ is the proximity measure. The computation of the value of $m(c_j,g)$ requires to query a given search engine in order to obtain the appropriate data. The hypothesis underlying this approach is that Wikipedia pages of close concepts are linked by a short path.

3.4 Web-Based Semantic Proximity Measures

Search engines are based on documents that are dynamically updated by a great number of users. Therefore using information on indexed terms provided by a search engine is then a valid approach to evaluate proximity semantics of pairs of terms, or groups of terms.

The general idea is to use a search engine or an object repository search tool S as a black box, to which submit queries and from which to extract useful statistics about the occurrence of a term or a set of terms in the indexed objects, as simple as counting the number of results for terms/objects queries. A proximity measure m is then used as heuristic $h(n)=m(n,g)$ to evaluate the most promising node n to browse, with the aim of reaching the goal node.[1] Since we normalize all the measures, distance and proximity measures can be compared as complementary.

- ***Confidence*** is an asymmetric statistical measure, used in rule mining for measuring trust in a rule $X \rightarrow Y$ i.e., given the number of transactions that contain X, then the *Confidence* in the rule indicates the percentage of transactions that contain also Y.

$$confidence(x \rightarrow y) = \frac{P(x,y)}{P(x)} = \frac{f(x,y)}{f(x)} \quad (2)$$

From a probabilistic point of view, confidence approximates the conditional probability, such that

$$confidence\ (x \rightarrow y) = P(y|x) \quad (3)$$

- ***Pointwise Mutual Information (PMI)*** [4] is a point-to-point measure of association used in statistics and information theory. Mutual information between two particular events w_1 and w_2, in this case two words w_1 and w_2 in webpages indexed by a search engine, is defined as:

$$PMI(w_1,w_2) = log_2 \frac{P(w_1,w_2)}{P(w_1)P(w_2)} \quad (4)$$

$PMI(w_1,w_2)$ is an approximate measure of how much a word gives information on the other word of the pair, in particular the quantity of information provided by the

[1] Let denote by $f(x)=S(x)$ and $f(x,y)=S(x \wedge y)$ the cardinalities of the results of a query to S, with search term x and $x \wedge y$ respectively, and we denote by N the number of documents/objects which are indexed by the search engine S. The probability estimate can be directly computed from the frequency, as $P(x)=f(x)/N$ (2), summarizing the frequency-based approach to probability, whenever the total N is known or can be realistically approximated with a value that will be greater than $f(x)$ for each possible x in the considered domain or context.

occurrence of the event w_2 about the occurrence of the event w_1. A high value of PMI represents a decrease of uncertainty. PMI has been successfully used in [15] to recognize synonyms, using only word counts, despite that on low frequency data, PMI does not provide reliable results and is not a good measure of independence, since values near zero indicate frequency, but at the same time is a bad measure of dependence, since the dependency score is related to the frequency of individual words.

- χ^2 **(Chi-squared or Chi-square)** [8] measures the significance of a relation between two categorical variables, checking if the values, observed by measuring frequency, differ significantly from the frequencies obtained by the theoretical distribution. The intuition behind Chi-square is that two events are associated only when they are more related than by pure chance. χ^2 coefficient has also been used in algorithms for community discovering.

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(a+c)(b+d)(c+d)} \quad (5)$$

where a,b,c,d represent the number of document in which w_1 and w_2 occurs according to $a= w_1 \wedge w_2$ $b= w_1 \wedge \neg w_2$ $c= w_2 \wedge \neg w_1$ $d=\neg w_1 \wedge \neg w_2$ and $n=N=a+b+c+d$.

- ***Normalized Google Distance (NGD)*** has been introduced in [3] as a measure of semantic relation, based on the assumption that similar concepts occur together in a large number of documents in the Web i.e., the frequency of documents returned by a query on a search engine S approximates the distance between related semantic concepts. The NGD between two terms x and y is formally defined as follows:

$$NGD(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - min\{\log f(x), \log f(y)\}} \quad \ldots\ldots\ldots\ldots \quad .(6)$$

where $f(x)$, $f(y)$ and $f(x,y)$ are the cardinalities of results returned by S for the query on x, y, $x \wedge y$ respectively, and M is the number of pages indexed by S, or a value which is reasonably greater than $f(x)$.

- ***PMING Distance*** [5][7][11] consists of NGD and PMI locally normalized, with a correction factor of weight ρ, which depends on the differential of NGD and PMI.

  More formally, the PMING distance of two terms x e y in a context W is defined, for $f(x) \geq f(y)$, as a function $PMING: W \times W \rightarrow [0,1]$:[2]

$$PMING(x,y) = \rho\left(1 - \log\frac{f(x,y)M}{f(x)f(y)\mu_1}\right) + (1-\rho)\left(\frac{\log f(x) - \log f(x,y)}{(\log M - \log f(y))\mu_2}\right) \ldots \quad \ldots(7)$$

[2] μ_1 and μ_2 are constant values which depend on the context of evaluation, defined as: μ_1=max PMI(x,y); μ_1=max NGD(x,y), with $x, y \in W$.

4 Experiments and Comparison

4.1 Experimental Setting

After a preliminary phase in order to evaluate general-purpose walk strategies and to tune the *WRW* strategy, systematic experiments on pairs of terms had been held on Bing as a search engine, with NGD and Confidence (CF) as heuristics, in order to evaluate the suitability of the different proximity measures as heuristics. Comparisons have been made among *HSW(BFS), HSW(Bing, CF, WRW)* and *HSW(Bing, NGD, WRW),* and ongoing work among *HSW(Wikipedia, CF, WRW)* and *HSW(Wikipedia, CF, WRW)* and *HSW(Wikipedia, NGD, WRW)* using in any case *Wikipedia* as a network, where a shortest path between the initial and target goal is searched, and a list of candidate terms for expansion was generated using the anchor links to other Wikipedia articles present in the textual content of the wiki page.

4.2 Pre-filtering and Bounds

Since the network is large and highly connected with non-relevant links, a pre-processing information filtering phase has been necessary, in order to filter out users' biases and non-semantic outliers:

Div Content. Only the anchor links in the content of the article are evaluated. [6][11].The parsing of the page can be furthermore limited by considering only the main content HTML *div* element of the article and not any other Wikipedia div or box.

Maximum number of candidate links (sub-optimality). In order to prune the graph and reduce computing time, a threshold can be stated on the number of linked candidate node [7]. In fact in Wikipedia the first lines of text are more related to the essential definition of a concept, while the longer a page is, the less significant links are provided at the end of the page.

Hub Links filtering. Filtering links that may lead to a hub in the Wikipedia network, with loss of semantic quality of the path: e.g., categories such as years, centuries and millenniums, first names of person, et cetera, nearly connect all the Wikipedia pages, without carrying a relevant semantic value.

Blank pages elimination. Pruning of pages without anchor links in the main text of the article i.e., dead ends.

Depth limitation. it has been proven useful for search speed-up to establish a limitation of the depth of the search, i.e. a maximum number of steps, although it is compromising completeness, it is in practice is not introducing a limitation.

4.3 Results Comparisons

Experimental result analysis shows that HSW has better performances than non-informed search and Pure Random Walk. In particular, all the considered proximity

measures used as a heuristics leads on the average to better results than a pure random walk, while non-informed search algorithms quickly exhausted memory resources or do not terminated for non-trivial searches.

For the sake of clarity, the results for non-informed search algorithms and the other experimented proximity measures are not shown in table 1 and 2: we can summarize the omitted result by noting that PMI generally performed worse than NGD and in some cases even worse than the PR pure randomness. Furthermore, the PMING distance, tested in preliminary experiments, was better performing than PMI itself, but required excessive computational cost, due to the many Wiki queries needed for context calculation, and was then discarded for the systematic experiments.

Chi-squared (CHI) was the third performing measure on the average although far behind NGD and Confidence.

The Bing search engine has been chosen as source of statistics because the more popular Google was found in [4] to lead to poorer results from a semantic point of view, due the bias and inconsistencies observed in the returned, probably for user modelling and commercial purposes, which no longer represent the real statistics of the engine. Analysing the results, in fact, it was observed that Google's results, both with and without API, greatly differ, while in and among the other search engines the difference is much smaller. This issue is explained by Google support as the lack of some additive services with the use of the API, where the lacking services are the ones which Google provides with the use of personal information about browsing, through cookies and/or accounting.

E.g. submitting the query "franzoni" to Google and Bing, the following results were obtained the following occurrences:

	With API	***Without API***
Google	1840000	5460000
Bing	2220000	2750000

And for the concurrent query "franzoni milani":

	With API	***Without API***
Google	28100	321000
Bing	44200	44300

We then decided to create a script to submit the queries and read the results through page scraping, instead of using the API, to automate the process and at the same time to keep the same results that would be obtained with a manual submission. Page scraping was implemented simply putting the HTML content of the page with query results in a string variable and extracting the data about the number of results matching regular expressions.

Notice that the cardinality of the results of a query in a search engine can vary and give different results in different times, so manually submitting a bunch of data may not return the best results, because of the time gap between pairs of terms or sets of terms which results have to be compared. Both with and without API is therefore suggested to submit queries with an automated program, to shorten the time lapse.

4.3.1 Performance Comparison

The performance comparison has been held by considering the average *path length* found from source to the target node, and the number of cases in which the algorithm does not converge at all on the given bounds. The best results were obtained with *HSW(Bing,CF,WRW)* and second best *HSW(Bing,NGD,WRW)*, resulting generally much more performing of PRW, see Fig.2(a).

Using the *CF+Bing* heuristics, the average number of steps needed to converge to the target node and the number of non-convergence cases is on average 50% of those required for the PRW, while for NGD+Bing the reduction is up to 25%, where comparison are made on 100 runs for each pair. Table 2 shows the performance for 100 runs on the pairs ("computer", "software"), ("planet", "galaxy"), ("drug", "abuse") and ("student", "professor"). In the first and second columns, the starting and goal terms of the pair are shown. In columns from third to fifth, the average number of steps needed for RW+Bing, NGD+Bing and CF+Bing to converge to the solution is shown. It is worth to mention that the NGD on worst case performed even better than the RW on average. Ongoing experiments with *HSW(Wikipedia, CF/NGD,WRW)* confirm the trend.

Table 1. experiment on h(n)={NGD, confidence} for the pair ("arithmetic", "counting")

Step 1: arithmetic=>{"mathematics", "science", "business"}

	t1	*f(t1)*	*t2*	*f(t2)*	*f(t1,t2)*	*NGD(t1,t2)*	*confidence(t1->t2)*
n_1	mathematics	101000000	counting	38800000	2230000	0,5495821639	0,0220792079
n_2	science	435000000	counting	38800000	7030000	0,5945563276	0,0161609195
n_3	business	864000000	counting	38800000	8780000	0,6614232480	0,0101620370

Step 2: mathematics=>{"quantity", "structure", "space"}

	t1	*f(t1)*	*t2*	*f(t2)*	*f(t1,t2)*	*NGD(t1,t2)*	*confidence(t1->t2)*
n_1	quantity	21600000	counting	38800000	1730000	0,4133869026	0,0800925926
n_2	structure	80100000	counting	38800000	2560000	0,4962758938	0,0319600499
n_3	space	383000000	counting	38800000	6190000	0,5945477630	0,0161618799

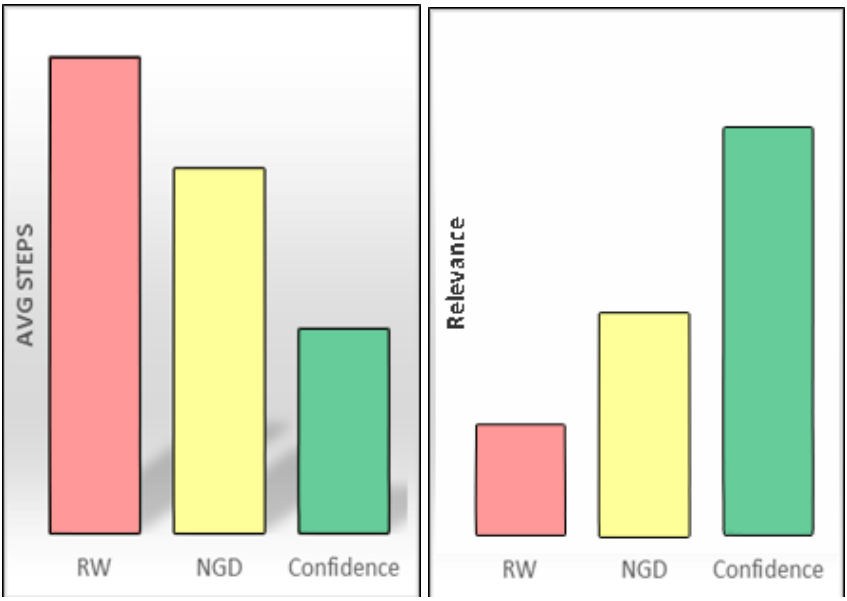
Step3: quantity=>{"property (philosophy)", "magnitude (mathematics)", "counting"}

	t1
n_1	property (philosphy)
n_2	magnitude (mathematics)
n_3	counting

HSW("arithmetic", "counting")={"arithmetic"→"mathematics" →"quantity" →"counting"}

Table 2. experiments on h(n)={NGD, PMI} on 100 runs for each pair

Start term	Goal term	Avg RW	Avg NGD	Avg CF
Computer	Software	159,48	91,1	19,42
Planet	Galaxy	94,66	48,52	17,94
Drug	Abuse	418,7	326,7	152,82
Student	Professor	225,16	184,6	153,21



(a: Performance) *(a: Human Evaluation)*

Fig. 2. Walk trends for Pure Random RW, NGD and Confidence)

4.3.2 Semantic Path Quality

Although a more systematic quality evaluation is an ongoing work, a preliminary evaluation confirms the best results of NGD and Confidence also with respect to the quality of the semantic paths returned, assessed by a blind human evaluation (see Fig.2(b) and Table 3) on a group of 15 users and 97 pairs (100 runs for Rt-HSW each pair). All the HSW variants returns a higher quality path, from a semantic point of view, than the PRW. The best quality resulted in HSW(Bing, NGD, WRW) as the best quality evaluation and HSW(Bing, Confidence, WRW) as the second best while the others navigate through less relevant intermediate nodes. This preliminary evaluation allows to draw the conclusion that HSW path is suitable to be used as a semantic explanation chain, for natural language contexts. The best quality and performance of Confidence over the other measures can be explained with the nature of the Wikipedia network, which contains articles explaining the meaning of terms, and links, which usually point to related explanatory information i.e., a directed graph, where links means explanation.

We argue that Confidence is more effective because is measuring more a casual relationships than a mere co-occurrence, like for example with NGD. The point is that Confidence is expected to be more suitable effective in order to navigate Wikipedia, since casual explanation is intuitively bringing higher quality to the semantic path. Furthermore, we have to consider, for instance, that more general concepts tend to have more inbound links, which reflects exactly the notion underlying Confidence; this pushes the search to go through more general concepts with respect to just co-occurring ones, then a semantic chain that is more easy to follow.

Table 3. Human evaluation (Avg) on heuristic-driven randomized strategies

Rt-HSW(Random Tournament)	**Relevance human evaluation (points from 0 to 5)**
PRW	1
RT-Confidence	4
RT-NGD	2

5 Conclusions

In order to solve the problem of extracting semantic paths between concepts from the Wikipedia collaborative network, we have proposed the Heuristic Semantic Walk approach, which uses search engine-based proximity measures as search heuristics. A remarkable feature of HSW is than it can be used for online search on large size semantic collaborative networks like Wikipedia, and that the HSW framework can be parameterized with respect to proximity measures and to the sources of information used for computing it. A proximity measure reflects some relationships between terms embedded in the indexed corpora of documents. The statistics extracted from search engines or from social networks reflect the relationships between terms and semantics as seen by the members of the specific community or social network. Experiments have shown that HSW with *Confidence* and *NGD* heuristics using *Bing* as a statistic source for semantic path searching in Wikipedia, outperforms classical search and other proximity measures, both with respect to *path length* and *quality*.

Ongoing research regards experimenting different proximity measures and variants of other informed search algorithms for semantic networks exploration, as well as systematic evaluation of semantic path quality. Future research will focus on further applications of the basic principle behind the HSW i.e., *using a semantic based heuristic to drive semantic search* to other contexts, such as, modelling user navigation in information repositories, modelling user associative reasoning in the area of natural language understanding and brain informatics.

References

1. Bollegala, D., Matsuo, Y., Ishizukain, M.: A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering (2011)
2. Cilibrasi, R., Vitanyi, P.: The Google Similarity Distance. ArXiv.org (2004)
3. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: ACL, vol. 27 (1989)

4. Franzoni, V., Milani, A.: PMING Distance: A Collaborative Semantic Proximity Measure. In: WI-IAT, 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 442–449 (2012)
5. Kurant, M., Markopoulou, A., Thiran, P.: On the bias of BSF. ITC (2010)
6. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. WIKIAI (2008)
7. Yeh, E., Ramage, D., Manning, C.D., Agirre, E., Soroa, A.: WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In: Proc. Graph-based Methods for Natural Language Processing (2009)
8. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. University of Michigan, MI (2003)
9. Cao, G., Gao, J., Nie, J.Y., Bai, J.: Extending query translation to cross-language query expansion with markov chain models. CIKM, ATM (2007)
10. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
11. Xu, Z., Luo, X., Yu, J., Xu, W.: Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency Computat: PE 23 (2011)
12. Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr Distance. Microsoft Research Asia (2008)
13. Leung, C.H.C., Li, Y., Milani, A., Franzoni, V.: Collective Evolutionary Concept Distance Based Query Expansion for Effective Web Document Retrieval. In: Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., Taniar, D., Apduhan, B.O., Gervasi, O. (eds.) ICCSA 2013, Part IV. LNCS, vol. 7974, pp. 657–672. Springer, Heidelberg (2013)
14. Gori, M.,, P.: A random-walk based scoring algorithm with application to recommender systems for large-scale e-commerce. In: 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)
15. Franzoni, V., Milani, A.: Heuristic Semantic Walk. In: Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., Taniar, D., Apduhan, B.O., Gervasi, O. (eds.) ICCSA 2013, Part IV. LNCS, vol. 7974, pp. 643–656. Springer, Heidelberg (2013)
16. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comp. Com. App (2006)
17. Franzoni, V., Milani, A.: Heuristic semantic walk for concept chaining in collaborative networks. International Journal of Web Information Systems 10(1), 85–103 (2014), doi:10.1108/IJWIS-11-2013-0031
18. Franzoni, V., Milani, A., Mengoni, P., Mencacci, M.: Semantic Heuristic Search in Collaborative Networks: Measures and Contexts. In: WI-IAT, 2014 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (2014) (accepted for)
19. Cheng, V.C., Leung, C.H.C., Liu, J., Milani, A.: Probabilistic Aspect Mining Model for Drug Reviews. IEEE Transactions on Knowledge and Data Engineering 99, 1 (preprint, 2014), doi:10.1109/TKDE.2013.175
20. Milani, A., Santucci, V.: Community of scientist optimization: An autonomy oriented approach to distributed optimization. AI Commun. 25(2), 157–172 (2012), doi:10.3233/AIC-2012-0526
21. Leung, C.H.C., Chan, A.W.S., Milani, A., Liu, J., Li, Y.: Intelligent Social Media Indexing and Sharing Using an Adaptive Indexing Search Engine. ACM TIST 3(3), 47 (2012), doi:10.1145/2168752.2168761
22. Baioletti, M., Milani, A., Poggioni, V., Rossi, F.: Experimental evaluation of pheromone models in ACOPlan. Ann. Math. Artif. Intell. 62(3-4), 187–217 (2011), doi:10.1007/s10472-011-9265-7