# Heuristic Semantic Walk

## Browsing a Collaborative Network with a Search Engine-Based Heuristic

Valentina Franzoni and Alfredo Milani

Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
{valentina.franzoni,milani}@dmi.unipg.it

**Abstract.** Path search between concepts over a semantic network is an issue of great interest for many applications, such as explanation generation and query expansion. In this study a new approach is proposed, to guide navigation over a collaborative concept network, in order to discover path between concepts. The method uses a semantic heuristic based on proximity measures, which reflects the collective knowledge embedded in search engines. The experiments held on the Wikipedia network and Bing search engine on a range of different semantic measures show that the proposed approach outperforms state of the art search methods.

**Keywords:** heuristics, information extraction, query expansion, semantic similarity measures, collaborative network, data mining.

## 1 Introduction

Many ways have been established to browse and search over a semantic network and several online semantic networks are available to the mass with interesting statistical information (e.g. search engines, social networks, collaborative encyclopaedias, et cetera). The problem of finding the semantic path between two or more given subjects in a semantic network, e.g. the path between two friends in a social network or the path between two entries of Wikipedia, is of great importance since it can provide useful information, concerning the subjects such as relationships and explanation. The intermediate subjects in the path over the semantic network can represent *hidden concepts* or *underlying concepts* which are implicit in the context of the two initially given concepts. Determining contextually hidden and implicit but relevant concepts is useful for a number of applications, such as:

*Automatic Explanation.* Let consider for instance two Wikipedia entries like *Ferrari* and *leather* and let the path *Ferrari* →*luxury cars* →*leather seats* →*leather*. The intermediate concepts in the chain can be used to generate an explanation of the relationship between *Ferrari* and *leather*.

*Natural language understanding.* The intermediate concepts in the chain can be used to provide a better context in order to disambiguate the meaning of some ambiguous terms or sentences used in natural language dialogs, by focusing on

B. Murgante et al. (Eds.): ICCSA 2013, Part IV, LNCS 7974, pp. 643–656, 2013.
© Springer-Verlag Berlin Heidelberg 2013

meanings which are consistent with the underlying context. For instance a speaker, hearing the word *seat* in a context *Ferrari, leather* will most probably link the concept of seat to a *leather seat in a luxury car* rather than to *wooden seat in a kitchen*.

*Query expansion*. The purpose of a query expansion system is to add to the search terms of a query a set of suitable keywords, in order to retrieve objects which have not been explicitly indexed under the original search terms. For instance, a query consisting of the two terms *Ferrari* and *leather* could likely be expanded using the *underlying keywords*: *luxury cars, leather seats* thus returning documents or pictures indexed under those additional keywords.

Collaborative semantic networks seem to represent the best available network on which to conduct a meaningful semantic search for some important reasons:

- *knowledge sharing oriented purpose*, in collaborative semantic network information are filtered and linked collaboratively, with the explicit purpose of knowledge sharing; for instance, in comparison to the links in a normal page of a web site, the links from a concept in Wikipedia to another one have a stronger relationship since they are purposely designed in order to provide explanations of concepts;
- *dynamic expansion and update*, a collaborative concept network is dynamically expanded and continuously refined by a multitude of users, whose aim is to improve the quality of the information and to share new information as the interest on those information arises; on the other hand artificially crafted semantic networks, such as the WordNet ontology, since they have a smaller number of authors, tend to reflect the biases of their authors and to be more static.

Our goal is to build a semantic chain of evidence that links two concepts in a semantic collaborative network (e.g. *Wikipedia* [10][14]). The problem of the semantic chain search can be reduced to a problem of search in a graph and can be applied, among others, to query expansion and building explanations. To establish which concepts are implied by a pair of terms in a dialogue in natural language, the path between the terms can be considered, where the starting and ending nodes in the path form the context.

In order to implement the search in the semantic chain, a pair of terms (*t1*, *t2*, i.e.: *start node*, *goal node*) is given, where the terms can be single words or textual expressions which correspond to Wikipedia articles. Starting from *t1*, the goal is to reach *t2* following the links of the network (i.e. the edges in the graph). In order to reach the goal node from the start node, an uninformed blind strategy can be applied, e.g. Breadth-First Search (BFS), Depth-First Search (DFS), Iterative Deepening Search (IDS) et cetera, because no additional information is given about the network.

In this study a new methodology is proposed, called *Heuristic Semantic Walk (HSW)*. In HSW a proximity measure (e.g. *confidence*, *Pointwise Mutual Information*, *Normalized Google Distance*) between concepts, derived from the statistical results of a query in a search engine, is used as *heuristic*, and applied to a walk to guide a path search over a collaborative concept network. The semantic heuristic, based on proximity measures, reflects the collective knowledge embedded in a search engine. The distance from each candidate successor of the current node *n* to the goal node *t2* is then calculated and a random tournament is exploited among all the distances *h(n)*, where randomness guarantees the completeness of the algorithm.

The experiments, held on the Wikipedia network on a range of different semantic proximity measures, show that the proposed approach outperforms uninformed search methods.

In particular, HSW returns the path which connects two concept nodes in much faster times than an uninformed blind search; 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. Different proximity measures which can be used as heuristics are discussed in section 3. The experimental results are then presented in section 4. Conclusions are drawn and future directions of the research are finally discussed.

## 2 The HSW Model

The problem we consider is to browse a semantic network in order to connect a pair of concepts, i.e. to search paths between nodes over an oriented graph $G = (V, E)$, where $V$ is a set of vertices/concepts (e.g. the entries in Wikipedia), and $E$ is a set of 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).

Several ways to browse a network are known, e.g. the blind random walk, [12][14][17] uninformed classical BFS, DFS algorithms, and their variants, or the several informed algorithms, such as the well known $A^*$ and its derivatives. Two main issues arise in the case of collaborative web-based semantic networks, such as Wikipedia: the graph dimension can be very large and dynamically changing; moreover is not clear what heuristic can be used for informed search algorithms.

The main idea of this study is to use a semantic proximity measure as heuristic, calculated from data extracted from collaborative collective sources of information such as general purpose search engines (e-g- Google, Bing) or specialized media repositories (e.g. YouTube, Flickr) or social networks (e.g. Facebook, Twitter).

### 2.1 HSW: The Problem

The goal of a HSW is to return the path between the pair of terms $(s, g)$, following the anchor links from the text of a starting Wikipedia article $s$, which corresponds to the first term of the pair, and driving the search towards the best successor candidate ($c_i$) using a semantic proximity measure, to a goal node ($g$), i.e. the corresponding Wikipedia article. The problem is to visit online the related Wikipedia pages, browsing from an article to the other through the anchor links in the article content, and returning as output the path chain and the number of steps (i.e. the path length).

### 2.2 Heuristic Based Search Strategy

The branching factor in Wikipedia can be very high (e.g. for the page “Rome” more than 500 links are present), so the BFS approaches are deeply penalized even in the

case in which the input terms are only moderately distant. On the other hand, Depth First approaches can not avoid to fall into loops, and Iterative Deepening Search, although complete, is extremely inefficient with high branching factors.

On the contrary, the Heuristic Semantic Walk is an informed search strategy, which makes use of a heuristic to estimate a score of each candidate node, to sort them for the expansion. The evaluation is performed in terms of closeness to the goal, where closeness is computed by measuring the proximity of the current concept to the goal.

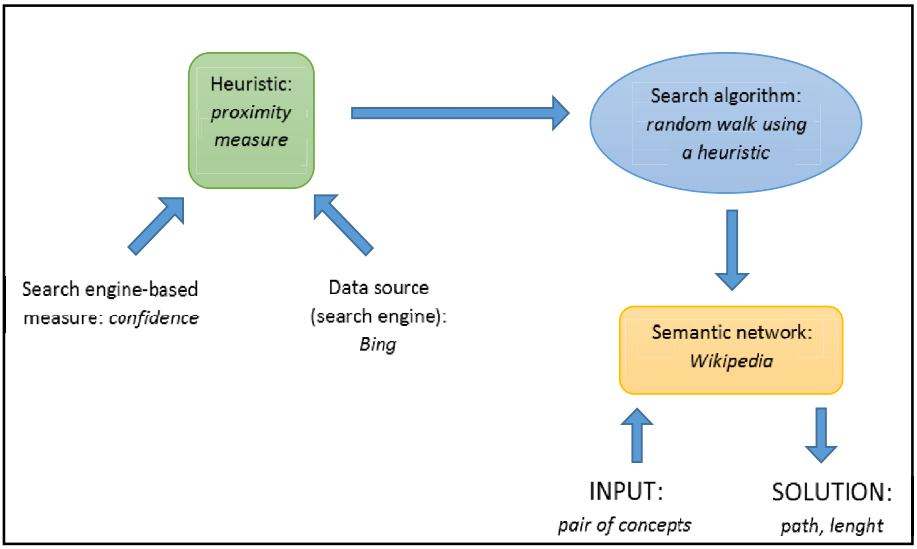


**Fig. 1.** Architecture of the Semantic Heuristic Walk

Another interesting point of HSW is that it allows an *online* search, i.e. the search agent can browse the Internet directly on the real Wikipedia, without constructing an offline graph with the Wikipedia content, and can evaluate the proximity measure online by directly querying the search engine in real time. The agent will open the links, read the related page, and extract the list of successor candidates by parsing the HTML, in other words it alternates a phase of *information gathering* with a phase of *exploration*, until the goal is reached.

### 2.3 HSW Exploration

The basic idea of the semantic heuristic driven walk is to generate a set of successors of the current node and to make a random tournament among the candidate successors using the probability distribution induced on the candidates by values of the proximity measure, between the candidate concepts and the target concept. 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

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

where $g$ is the target node, $c_j$ are the successors of the current node *curr*, and $Prox(c_j, g)$ is the proximity measure.

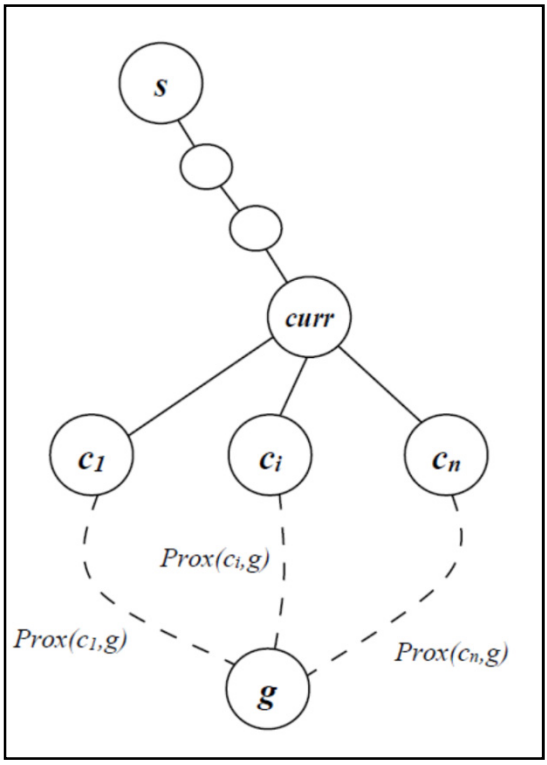


**Fig. 2.** Random tournament in HSW

Figure 2 shows the decision stage of HSW from a starting concept *s* to a target concept *g*. The computation of the value of $Prox(c_j, g)$ require to query a given search engine (see Figure 1) in order to obtain the appropriate data.

The hypothesis underlying this approach is that Wikipedia pages of close concepts are close, i.e. have a short path linking them, therefore the use of a suitable proximity measure which reflects the human judgement of closeness/relatedness is important in order to guarantee the adequacy of the heuristic.

### 2.4 HSW Information Gathering

Since the Wikipedia network has a general structure, i.e. does not exhibit a hierarchical structure like taxonomy networks, the HSW is prone to follow redundant paths and to diverge temporarily from the expected semantic result.

The phase of information gathering consists in the analysis of the content of a Wikipedia page by parsing it and doing preliminary filtering operations. Some of these operations, such as link extraction, can be very much resource consuming; filtering is also an important operation since it can avoid searching along useless or trivial/meaningless path in the network.

In order to optimise the semantic path search, some context-driven optimisation strategies can be applied in the Wikipedia domain:

1. Div limitation
2. Maximum number of candidate links (suboptimality)
3. Link filtering
4. Blank pages elimination
5. Depth limitation

In particular:

*Div limitation*. Since the focus is on the explanation of the concepts, only the anchor links in the content of the article are evaluated. [10] The parsing can be furthermore limited by considering only the main content HTML *div* element of the article and not any other Wikipedia div or box. In our experiments, this strategy lead to results with a higher semantic quality, with respect to natural language and human evaluation.

*Maximum number of candidate links*. To prune the graph and reduce computing time, a threshold can be stated on the number of candidates. In theory this step is at the expense of optimality, but in practice pruning the dictionary cutting off the candidate links in a Wikipedia graph was found to lead to higher results, [11] since one article is more general when the number of *inlinks* is larger, and smaller graphs and dictionaries increase the quality. [12] Our consideration is that 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. Therefore, according also to preliminary tests conducted in this study, giving a threshold on the number of links will not produce a loss of information, but on the contrary can lead to higher semantic quality results.

*Link filtering*. In order to prune the dictionary, a further optimization can be performed, filtering some specific kinds of links that may lead to a hub in the Wikipedia network, with loss of semantics: e.g. categories such as years, centuries and millenniums, first names of person, etc., nearly connect all the Wikipedia pages but they do not carry a semantic value; they are useful to the user for the purpose of quick information retrieval like alphabetical indexes, but they do not represent relevant semantic relationships.

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

*Depth limitation*. For very quick searches, to state a limitation of the depth of the search can be useful, i.e. a maximum number of steps, at the expense of completeness.

## 3 Search Engine-Based Semantic Proximity Measures

As noted before, search engines are the natural source of semantic information. Search engines are based on documents which are dynamically updated by a great number of users, then using information on indexed terms provided by a search engine is a valid approach to evaluate proximity semantics of pairs of terms, or groups of terms.

The general idea is to use search engines as a black box to which submit queries and extract useful statistics, to evaluate proximity semantics about the occurrence of a term or a set of terms, just counting the number of results for that terms. The proximity measure is then used as heuristic *h(n)* to evaluate the most promising node *n* to browse, with the aim of reaching the goal node.

Let define *f(x)*, *f(x,y)* respectively as the cardinalities of the results of a query of term *x* and *xANDy*, and *N* the number of documents which are indexed by the search engine.

The probability is directly deductible from the frequency:

$$P(x) = \frac{f(x)}{N} \tag{2}$$

(2) summarizes the frequency based approach to probability, i.e. in the following formula, probability *P* can be computed from frequency *f* and vice versa, 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.

### 3.1 Confidence and Average Confidence (CM)

Given a rule $X \rightarrow Y$, confidence is a statistical measure that, given the number of transactions which contain *X*, indicates the percentage of transactions which contain also *Y*. Confidence is a symmetric function.

$$confidence(x \rightarrow y) = \frac{P(x,y)}{P(x)} = \frac{f(x,y)}{f(x)} \tag{3}$$

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

$$confidence(x \rightarrow y) = \frac{\mathrm{P}(xy)}{\mathrm{P}(x)} = \mathrm{P}(y|x) \tag{4}$$

Average Confidence (CM) can be defined as

$$CM = \frac{confidence(\mathrm{x}\rightarrow\mathrm{y}) + confidence(\mathrm{y}\rightarrow\mathrm{x})}{2} \tag{5}$$

### 3.2 Pointwise Mutual Information (PMI)

*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 the occurrence of particular words in Web-based text pages, is defined as

$$PMI(w_1, w_2) = log_2 \frac{P(w_1,w_2)}{P(w_1)P(w_2)} \quad (6)$$

This type of mutual information 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 occurrence of the event $w_2$ about the occurrence of the event $w_1$, i.e. the conditional probability $w_1|w_2$. This measure is used in classical information retrieval considering the position of words in a textual corpus, to estimate how much the likelihood of having the word $w_1$ in position $i+1$ will increase if we have $w_2$ in position $i$. A high value of PMI represents a decrease of uncertainty.

The PMI has been successfully used in [19] to recognize synonyms, using only the count of words on the Web. On particularly low frequency data, PMI does not provide reliable results. Since PMI is a ratio of the probability of $w_1$, $w_2$ together and $w_1$, $w_2$ separately, just consider the two extreme cases: the case of a perfect dependence of occurrences of both words (i.e. they occur only together) and perfect independence (i.e. they never occur together).

In the case of perfect dependence, PMI will be:

$$PMI(w_1, w_2) = log_2 \frac{1}{P(w_2)}$$

In the case of perfect independence, PMI will be:

$$PMI(w_1, w_2) = log_2 \frac{P(w_1)P(w_2)}{P(w_1)P(w_2)} = log_2 1 = 0$$

We can therefore say that PMI is 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. In addition, pairs of terms with low frequency will receive a greater score than pairs of terms with high frequency, so PMI could not always be suitable when the aim is to compare information on different pairs of words.

### 3.3 Chi-square Coefficient ($\chi^2$)

$\chi^2$ (*Chi-squared* or *Chi-square*) makes possible to assess 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.

In common parlance, two events are associated where you can define a relationship between them, but in statistics two events are associated only when they are more related than by pure chance.

The question to which Chi-square can answer is "how much the observed data deviate from those that would be expected if they were random?".

For each value, consider the quantity:

$$\frac{(observedvalue - expectedvalue)^2}{expectedvalue}$$

where the numerator is squared to always get a positive number, even when the expected value is greater than the observed value. It is evident that this quantity increases when the difference between the compared data increases, i.e. when the data can be considered significantly different from randomness. The sum of that amount on two values enables to calculate the relative significance of their co-occurrence: the higher the $\chi^2$, the greater the likelihood that the relation is not random, and therefore significant.

Given two events $W_1$ and $W_2$, in this case the occurrence of particular words in Web-based text pages, let define:

- $a = W_1 \wedge W_2$ (number of documents where $W_1$ and $W_2$ occur);
- $b = W_1 \wedge \neg W_2$ (number of documents where $W_1$ occurs, but not $W_2$);
- $c = W_2 \wedge \neg W_1$ (number of documents where $W_2$ occurs, but not $W_1$);
- $d = \neg W_1 \wedge \neg W_2$ (number of documents where neither $W_1$ nor $W_2$ occur);
- $n = N = a + b + c + d$

An algebraically simplified formula to calculate Chi-square is the following: [9]

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(a+c)(b+d)(c+d)} \tag{7}$$

where the coefficient of association can be directly calculated from the observed data, without having to calculate the related expected values.

The $\chi^2$ coefficient has also been used in community discovering algorithms. [15]

### 3.4 Normalized Google Distance (NGD)

In 2006 the *Normalized Google Distance (NGD)* [3] was presented 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. that the frequency of documents returned by a query on Google or any other search engine approximates the distance between related semantic concepts. Notice that the NGD was originally defined for Google, but it is a measure that can be applied to any search engine, so "NGD" is not the same of "distance on Google".

The NGD between two terms *x* and *y* is formally defined as follows:

$$NGD(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - min\{\log f(x), \log f(y)\}} \tag{8}$$

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

If the two terms *x* and *y* do not ever occur in the same document, but occur separately (i.e. *x* and *y* are not in relation), their NGD should be *1*; otherwise, if *x* and *y* always co-occur (i.e. *x* and *y* are identical), their NGD should be *0*.

The NGD is not a metric. In fact it does not respect the property for which *NGD(x,y)>0* for each *x≠y*, because it can be possible that two terms *x* e *y* have the same cardinality of results:

$$\mathrm{f(x) = f(y) = f(x,y) \rightarrow NGD(x,y) = 0}$$

At the same time it does not respect also the property of triangular inequality:

$$\mathrm{NGD(x,y) <= NGD(x,z) + NGD(z,y)} \quad \text{for each } x, y, z.$$

E.g. Given

$$z=(xORy),\ (x\ AND\ y)=0,\ x=xANDz,\ y=yANDz,\ f(x)=f(y)=\sqrt{N}$$

then

$$f(x,z)=f(z,y)=\sqrt{N},\ f(z)=2\sqrt{N} \text{ and } f(x,y)=0$$

so that *NGD(x,y)=1*, *NGD(x,z)=NGD(z,y)=2/log(N/4)*, where *1>4/log(N/4)* (QED).

Although it is not a metric, NGD is a good measure of proximity, which turns in a variety of experimental applications.

E.g. Consider the three terms "*saturn*", "*donkey*" and "*horses*" and construct the matrix of their co-occurrences in Google. Calculating the NGD for pairs "*saturn-donkey*" and "*donkey-horses*", it will return a higher value for "*saturn-donkey*" than for "*donkey-horses*": it means that the semantic distance for the pair "*saturn-donkey*" is greater than that for "*donkey-horses*", and this does not betray common sense.

### 3.5 PMING Distance

*PMING Distance* [5] 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)≥f(y)*, as a function *PMING:W×W→[0,1]*:

$$PMING(x,y) = \rho\left(1 - \log\frac{f(x,y)M}{f(x)f(y)\mu_1}\right) + (1-\rho)\left(\frac{\log f(x) - \log f(x,y)}{(\log M - log f(y))\mu_2}\right) \qquad (9)$$

where:

- $\rho$ is a parameter to balance the weight of components;
- $\mu_1$ e $\mu_2$ are constant values which depend on the context of evaluation, and are defined as:

$$\mu_1 = \max \mathrm{PMI}(x,y), \text{ with } x, y \in \mathrm{W}$$

$$\mu_2 = \max \mathrm{NGD}(x,y), \text{ with } x, y \in \mathrm{W}.$$

The PMING Distance has its main application in the use with the main search engines as source of information about the occurrence of terms or sets of terms (i.e. the

cardinality of the set of documents returned as result) and can be applied to every object that can be measured with frequency or probability in a text corpus.

## 4 Experimental Results

Experiments were conducted on sample pairs of terms, both using the DLS and the HSW. The input terms can be words or textual expressions, with the only constraint to be related to a Wikipedia page.

Confidence and NGD were used as heuristics and Bing was used as a data source to evaluate the proximity of the candidate terms. Wikipedia was used as a semantic network and dictionary where to extract and browse the initial, final and candidate concepts. In this context, the input pair is related to Wikipedia pages, and the candidate terms list is generated with the anchor links to Wikipedia articles, present in the content text of the page.

For each of the pairs *(t1, t2)* of Wikipedia articles given as input, the first 3 anchor links in the content DIV of the Wikipedia article corresponding to the first term are stored in a vector $v=(n_1, n_2, n_3)$, filtering the words belonging to general categories, such as years, months, disambiguation hubs, et cetera. For each term $n_i$ in the vector, with $i=1,...,3$ (i.e. for each candidate term for the expansion), the heuristic $h(n_i)$ is calculated as the distance between the term itself and the second term of the input pair (i.e. the goal term *t2*). A weighted tournament is exploited to choose one node in *v*, evaluating the score of each node on the basis of the heuristic function. The chosen node will be then submitted as *t1* for the next step.

**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"}

E.g. Table 1 shows the results for the HSW of the pair ("arithmetic", "counting"): *HSW("arithmetic", "counting") ={"arithmetic" → "mathematics" → "quantity" → → "counting"}*

The HSW returns a higher quality path, in a semantic point of view, than the classical blind search. The experiments show that the HSW path is suitable to be used as a semantic explanation chain, for a natural language context in which the input pair is included.

## 5 Conclusions

Searching paths between concepts in a semantic network is interesting for knowledge based applications which can use the path found as a base for *query expansion*, *explanations generation* and *hidden contextual knowledge*. Except for the case of structured semantic networks, such as taxonomies, efficient searching in huge semantic networks is open issue.

The Heuristic Semantic Walk model approaches the problem of searching paths in a semantic network by using heuristics based on proximity measures between the candidate terms and the goal term, the proposed heuristics can be computed from data obtained by querying a search engine. The HSW model has been experimented by searching paths on the Wikipedia network, and using the PMING Distance as proximity measure evaluated on the Bing search engine.

Compared to previous approaches proposed to explore semantic networks HSW is the first approach, to the best of our knowledge, which uses a search engine-based proximity measure as heuristic. Another remarkable feature of HSW is than it can be performed on online networks, such as the web based huge semantic network of Wikipedia, without the need of constructing a separate offline network.

Another important element of HSW is that the search strategy can be contextualized depending on which proximity measure and which source of information are used for computing it. A proximity measure reflects the relationships between terms embedded in the indexed corpora of documents. In this way data extracted from specialized search engines (e.g. Flickr or YouTube), or from social networks (e.g. from messages, chats, forum etc.) will more appropriately reflect the relationships between terms as seen by the members of the specialized community or social network.

Ongoing research regards experimenting different proximity measures and derivatives of informed search algorithms, like A*, in order to determine the optimal choice for exploring specific semantic networks.

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*. Examples of these kind of applicative contexts are for instance: modelling users' navigation in information repositories (e.g. applications which support the user exploring a website), modelling users' associative reasoning (e.g. in applications in the field of natural language understanding and brain informatics).

**Acknowledgements.** The authors thank Marco Mencacci and Paolo Mengoni, students of the Computer Science Master degree, University of Perugia, whose *WikiDist* Python project, developed for the course of Artificial intelligence, was successfully used in this study.

## References

1. Etzioni, O.: Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web. AAAI (1996)
2. 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)

3. Cilibrasi, R., Vitanyi, P.: The Google Similarity Distance. ArXiv.org (2004)
4. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: ACL, vol. 27 (1989)
5. Franzoni, V., Milani, A.: PMING Distance: A Collaborative Semantic Proximity Measure. In: WI-IAT, vol. 2, pp. 442–449 (2012); IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (2012) ISBN: 978-1-4673-6057-9, doi:10.1109/WI-IAT.2012.226
6. Kurant, M., Markopoulou, A., Thiran, P.: On the bias of BSF. ITC (2010)
7. Franzoni, V.: Semantic Proxymity Measures for the Web (Misure di Prossimità Semantic per il Web), Laurea Thesis, Department of Mathematics and Computer Science, Università degli Studi di Perugia, Italy (2012)
8. Franzoni, V., Gervasi, O.: Guidelines for Web Usability and Accessibility on the Nintendo Wii. In: Gavrilova, M.L., Tan, C.J.K. (eds.) Transactions on Computational Science VI. LNCS, vol. 5730, pp. 19–40. Springer, Heidelberg (2009)
9. Manning, D., Schutze, H.: Foundations of statistical natural language processing. The MIT Press, London (2002)
10. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: WIKIAI (2008)
11. Gabrilovich, E., Markovich, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: IJCAI (2007)
12. 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)
13. Mukhopadhyay, D., Banik, A., Mukherjee, S., Bhattacharya, J., Kimin, Y.: A Domain Specific Ontology Based Semantic Web. Search Engine. Distributed Computing (2011)
14. Smith, J.R., Quirk, C., Toutanova, K.: Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010)
15. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. University of Michigan, MI (2003)
16. Richards, I.A., Ogden, C.K.: The Meaning of Meaning. A Study of the Influence of Language upon Thought and of the Science of Symbolism (1923)
17. Cao, G., Gao, J., Nie, J.Y., Bai, J.: Extending query translation to cross-language query expansion with markov chain models. In: CIKM. ATM (2007)
18. Cialdea Mayer, M., Limongelli, C., Orlandini, A., Poggioni, V.: Linear temporal logic as an executable semantics for planning languages. Journal of Logic, Language and Information 16(1) (2007)
19. 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)
20. Xu, Z., Luo, X., Yu, J., Xu, W.: Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency Computat.: Pract. Exper. 23 (2011)
21. Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr Distance. Microsoft Research Asia (2008)
22. Leung, C.H.C., Chan, W.S., Milani, A., Liu, J., Li, Y.X.: Intelligent Social Media Indexing and Sharing Using an Adaptive Indexing Search Engine. ACM Transactions on Intelligent Systems and Technology (2012)

23. Santucci, V., Milani, A.: Particle Swarm Optimization in the EDAs framework. In: Gaspar-Cunha, A., Takahashi, R., Schaefer, G., Costa, L. (eds.) Soft Computing in Industrial Applications. AISC, vol. 96, pp. 87–96. Springer, Heidelberg (2011)
24. Santucci, V., Milani, A.: Community of Scientist Optimization An autonomy oriented approach to distributed optimization. AI Communications 25(2), 157–172 (2012) ISSN (print): 0921-7126, ISSN (online): 1875-8452, doi: 10.3233/AIC-2012-0526
25. Santucci, V., Milani, A.: Adaptive Memetic Particle Swarm Optimization. In: Proceedings of 16th Online Conference on Soft Computing in Industrial Applications (WSC16)
26. Santucci, V., Milani, A.: Community of Scientist Optimization: Foraging and Competing for Research Resources. In: IJCAI Workshop Proceedings, 18th RCRA International Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion, pp. 66–80 (2011)
27. Milani, A., Baioletti, M., Santucci, V.: Discrete Differential Evolution for Learning Bayesian Network Structure. In: Proceedings of GECCO, Genetic and Evolutionary Computation Conference (2013)
28. Milani, A., Santucci, V.: Particle Swarm Estimation of Distribution Algorithm for Lymphoma Classification through Automatic Biopsies Analysis. In: Proceedings of Mibisoc, International Conference on Medical Imaging using Bio-inspired and Soft-Computing (2013)
29. Milani, A., Ukey, N., Niyogi, R., Poggioni, V., Singh, K.: A Bidirectional Heuristic for Web Service Composition with Costs. International Journal of Web and Grid Services, Inderscience 6, 160–175 (2010)
30. Milani, A., Poggioni, V.: Planning in Reactive Environments. Computational Intelligence 23, 439–463 (2007)
31. Milani, A., Santucci, A.V., Leung, V.C.: Optimal Design of Web Information Contents for E-Commerce Applications. In: Gelenbe, E., Lent, R., Sakellari, G., Sacan, A., Toroslu, H., Yazici, A. (eds.) Computer and Information Sciences. LNEE, vol. 62, pp. 339–344. Springer, Heidelberg (2010)
32. Leung, C.H.C., Milani, A., Franzoni, V., Li, Y.X.: Collective Evolutionary Concept Distance Based Query Expansion for Effective Web Document Retrieval. LNCS (in press, 2013)
33. Gentili, E., Milani, A., Poggioni, V.: Data Summarization Model for User Action Log Files. In: Murgante, B., Gervasi, O., Misra, S., Nedjah, N., Rocha, A.M.A.C., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2012, Part III. LNCS, vol. 7335, pp. 539–549. Springer, Heidelberg (2012)