# Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts

**Robert West** and **Joelle Pineau** and **Doina Precup**
School of Computer Science
McGill University
Montréal, Québec, Canada
{rwest, jpineau, dprecup}@cs.mcgill.ca

## Abstract

Computing the semantic distance between real-world concepts is crucial for many intelligent applications. We present a novel method that leverages data from 'Wikispeedia', an online game played on Wikipedia; players have to reach an article from another, unrelated article, only by clicking links in the articles encountered. In order to automatically infer semantic distances between everyday concepts, our method effectively extracts the common sense displayed by humans during play, and is thus more desirable, from a cognitive point of view, than purely corpus-based methods. We show that our method significantly outperforms Latent Semantic Analysis in a psychometric evaluation of the quality of learned semantic distances.

## 1 Introduction

For a host of intelligent computer applications, commonsense knowledge is helpful or even essential, notably in fields such as automated reasoning, information retrieval and natural language processing. For instance, imagine an automated shopping agent crawling the Web for bargains on behalf of a human customer, e.g. trying to find a digital camera. It would be useful to know that a digital camera is a kind of electronic device, or at least that the concepts DIGITAL CAMERA[1] and ELECTRONICS are highly related, because the agent will encounter the specific expression DIGITAL CAMERA only rarely during its forage; but once it finds related words such as ELECTRONICS, it can push forward on those links, knowing it is likely to find something relevant there.

There have been numerous attempts to use hand-crafted semantic networks, most notably WordNet [Fellbaum, 1998], to infer this kind of relatedness information (see [Kaur and Hornof, 2005] for an overview). For the AI community, it seems more interesting to infer such semantic knowledge automatically from raw data, and thanks to the Internet, substantial amounts of useful data are readily available. In practice, however, a large part of the Web is not easily amenable to automated analysis because natural language understanding is still a very hard task. One can circumvent this problem by initially focusing on highly structured sub-Webs, such as Wikipedia. Wikipedia contains tremendous amounts of reliable text that could be fed to NLP software, but even its raw hyperlink structure carries a significant amount of interesting information. The hypertextual Wikipedia graph alone can be viewed as a very primitive semantic network: articles represent the concepts, while hyperlinks (sometimes) represent semantic relationships between the concepts they connect. With this in mind, it seems reasonable to use techniques developed to infer the degree of relatedness between two concepts based on their relative position in a semantic network using the Wikipedia graph instead of, say, WordNet. Rada *et al.* [1989], e.g., proposed a shortest-path metric, according to which the degree of relatedness is determined by the length of the shortest path between two vertices in the semantic network graph.

But using the raw hyperlink structure of Wikipedia leads to several problems. First, while many hyperlinks correspond to semantic links, many others do not. Links are often added based on the inclination of the author, rather than because the concepts are related. Also, if one looks only at the presence or absence of links, no distinction can be made between closely and loosely related concepts. This leads to a combinatorial explosion, such that every page is connected to every other page by 4.6 links on average [Dolan, undated]. For instance, both BASEBALL and ARCHIMEDES have distance 2 to CARL FRIEDRICH GAUSS according to the shortest-path metric, although clearly the latter is relevant while the former is not.

This 'small world' phenomenon is particularly problematic in Wikipedia, but it is also a concern in hand-crafted semantic networks like WordNet. One way to deal with it is to augment the shortest-path metric by accounting for the frequency of two concepts co-occurring in a text corpus [Jiang and Conrath, 1997]; simply put, the more often two concepts co-occur, the more heavily weighted the link between them should be. This approach incorporates human knowledge indirectly, by analyzing a collection of documents written by people. In this paper we present a novel way of exploiting human common sense more directly. We use data harvested from a word association game that is played on Wikipedia. Human players are asked to navigate between different articles, and in this process they provide click frequency information, effectively (and unwittingly) weighting links in the graph. Over several games, links that are more relevant semantically will end up being more heavily weighted. We

[1] We will use SMALL CAPS to denote concepts.

propose a novel algorithm for computing semantic distances from this data. The algorithm is inexpensive, but its results are cognitively plausible.

The method we propose is a major step towards transforming the Wikipedia hyperlink graph into a semantic network. Our approach filters irrelevant links, while keeping the relevant ones and (indirectly) adding missing ones. Weights for these links are computed from the click information. The only missing step to obtain a proper semantic network is labeling the links with the type of relationship they represent (e.g. 'is-a', 'is-part-of'), for which click information is insufficient.

The paper is structured as follows. In Sec. 2, we describe the game and the data we gather. Sec. 3 describes our semantic distance measure. In Sec. 4, we propose an approach for deciding if two concepts are sufficiently related or not. In Sec. 5, we present results, which suggest that our method performs better than Latent Semantic Analysis, a standard corpus-based approach. Finally, Sec. 6 discusses related work, and Sec. 7 contains conclusions and avenues for future work.

## 2 Game Description

The idea of designing games to collect data from humans has been championed by von Ahn and colleagues. For example, 'Verbosity' [von Ahn *et al.*, 2006] is similar to the popular 'Taboo' game and aims at collecting common-sense facts.

Our game, 'Wikispeedia', is a version of the 'Wiki Game' [Wikipedia, 2009] that has been played casually by Wikipedia users for a while. To the best of our knowledge, no analysis of data from this (or any similar) game has been done to date.

People play the game individually. The player is given two Wikipedia articles (or alternatively, he/she can choose them). Starting from the first article, the goal is to reach the second one (the *goal article*), exclusively by following links in the articles encountered, minimizing the number of link clicks. Step-by-step backtracking is possible 'for free'.

There is a crucial difference between the way a computer and a human would play this game. A computer would simply find the shortest path between the start and the goal, by any standard algorithm. This is clearly impractical for most humans. (A cheater could code a shortest-path finder, but we ignore this problem for now.) A human player will instead leverage semantic associations based on background knowledge of many common sense facts, and select links according to this knowledge. Consider the task of finding a path from SEYCHELLES to GREAT LAKES. It was solved in an actual game instance as follows: ⟨SEYCHELLES, FISHING, NORTH AMERICA, CANADA, GREAT LAKES⟩. This example showcases the anatomy of a typical game. Players try to reach, as quickly as possible, a general concept (in this case NORTH AMERICA), whose article has a lot of outgoing links. From such *hubs* it is easy to reach many parts of the Wikipedia graph. After this initial 'getting-away' phase, the 'homing-in' phase starts: the search narrows down again towards more specific articles that get more and more related to the goal.

There is a striking difference between the human path and the result of a shortest-path algorithm for this example: SEYCHELLES and GREAT LAKES are optimally connected by ⟨SEYCHELLES, ASIA, AMERICAN ENGLISH, GREAT LAKES⟩, which is far less semantically meaningful than the path found by the human. In general, humans find intuitive, not shortest, paths. This observation is corroborated by Fig. 1, which shows that the distribution of game path lengths is shifted towards longer paths, compared to the shortest-path algorithm, and that it has a heavy tail towards longer paths.

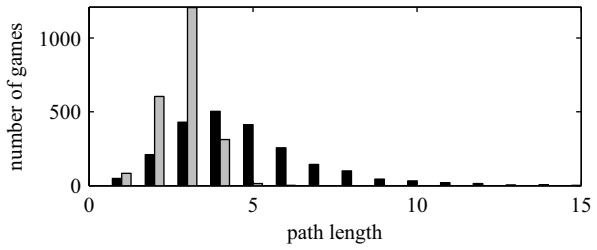


Figure 1: *Black:* histogram of lengths of 1,694 games; the tail continues up to 30. *Gray:* histogram of shortest-path solutions to the same games.

Given that people make heavy use of their common-sense knowledge in playing this game, it is desirable to develop a method able to extract this knowledge from recorded game traces. Note that, in general, the process of collecting knowledge from people is very time-consuming and thus expensive. For example, professional lexicographers had to be employed to build WordNet, and every time the database has to be extended, experts have to be consulted again. Our approach has a substantial advantage, in that we pay contributors with fun instead of money, hence obtaining the data is easier. Moreover, this approach is easily scalable: if we want to add a new concept, we simply make it the goal of some future games. The fact that the informal 'Wiki Game' has been popular among Wikipedians for a long time means that such adaptive data collection can be achieved quickly and at low cost.

We will now explain how we use the game traces to define a semantic distance measure between concepts.

## 3 Proposed Semantic Distance Measure

A purely path length–based measure could not account for the frequency with which players choose an article to reach a goal. But clearly, if many players pick a specific article, it should be considered more related to the goal than if only few do. This is why the semantic distance measure we propose is based on information theory. Intuitively, it quantifies how many bits are needed to encode a common-sense Wikipedia path between two concepts. The fewer bits are needed, the more strongly the two concepts are related. In order to formalize this idea, we must first discuss *click probabilities*.

Let $A$, $A'$ and $G$ be random variables representing the current Wikipedia page, the next Wikipedia page and the goal page of a game. For any Wikipedia article $a$ and any Wikipedia goal (or target) article $g$, one can consider the probability distribution $P(A'|A=a,G=g)$ over $a$'s out-links. This distribution is multinomial and specifies, for each article $a'$ that can be reached in one hop from $a$, the probability that a player continues to $a'$ if he/she is currently on $a$ and is trying to find goal article $g$. This can be estimated from the ob-

served games using standard Bayesian methods, as the mean of the Dirichlet distribution which is the conjugate prior of $P(A'|A=a,G=g)$. We use $P^*$ to denote the *posterior click probability* estimated after seeing all the data:

$$P^*(A'=a'|A=a,G=g)=\frac{N(A'=a',A=a,G=g)+\alpha}{N(A=a,G=g)+\alpha L_a}, \quad (1)$$

where $\alpha$ is the Dirichlet parameter representing the initial confidence in the uniform prior distribution, $L_a$ is $a$'s outdegree (i.e. the number of articles linked from $a$), $N(A=a,G=g)$ is the number of times $a$ was encountered on paths for which $g$ was the goal, and $N(A'=a',A=a,G=g)$ counts how often the link to $a'$ was chosen in this situation.

Before observing any games (i.e. if all $N$-counts in (1) are zero) the estimate is the uniform *prior click probability:*

$$P^0(A'=a'|A=a,G=g)=1/L_a \quad (2)$$

Now consider one particular path $p=\langle a_1,a_2,...,a_n=g\rangle$. We can compute a *path-specific distance* from every article $a_i$ along $p$ to the goal $g$, i.e. for every $i$ with $1\leq i<n$ we get

$$d_p(a_i,g)=\frac{-\sum_{j=i}^{n-1}\log P^*(A'=a_{j+1}|A=a_j,G=g)}{-\log \text{PageRank}(g)}. \quad (3)$$

In the numerator, $-\log P^*(A'=a_{j+1}|A=a_j,G=g)$ is the information content of the link from $a_j$ to $a_{j+1}$ given that the goal is $g$, or in other words, the number of bits needed to represent that link optimally in a Huffman coding. So the numerator sums up the numbers of bits needed to code each separate link that was clicked along $p$, and consequently indicates the number of bits needed to code the entire path (note that this is conditional on $g$).

The denominator contains the Google PageRank [Brin and Page, 1998] of the goal article $g$, which is the stationary probability of $g$ during a (fictional) random walk on the Wikipedia graph. We implemented the PageRank algorithm and ran it locally on the Wikipedia graph to get these numbers. One can think of PageRank($g$) as the prior probability of being in article $g$, and of the entire denominator as $g$'s information content, or the number of bits needed to code article $g$ independently of any game. This serves the purpose of normalization: intuitively, a concept that is hard to reach (hard to 'explain') is allowed to be related to concepts that are farther from it on Wikipedia paths. For instance, UNITED STATES has PageRank 0.010 (1% of time steps on a random walk will be spent on the UNITED STATES article), while TURQUOISE has a PageRank of only $5.8\times10^{-5}$. Since $-\log(0.010)\approx 6.6$ and $-\log(5.8\times10^{-5})\approx 14$ (about twice 6.6), a path from an article $a$ to goal TURQUOISE may take twice as many bits to code as a path from some article $b$ to goal UNITED STATES, and still we will have $d(a,\text{TURQUOISE})\approx d(b,\text{UNITED STATES})$.

Instead of using uniform transition probabilities for the random walk (cf. (2)), as in the standard PageRank algorithm, it might seem better to use the transition probabilities estimated from data (cf. (1)). Such a 'posterior PageRank' would indicate how hard it is to find an article while one is actively looking for it, rather than wandering aimlessly. Numerically, however, this is a minor difference, so the results we present here use the standard PageRank.

So far, we have described distances that are derived from single paths. To get a *path-independent distance* from $a$ to $g$, we simply average over all paths running through $a$ and reaching goal $g$. Thus, if there are $m$ such paths $p_1,p_2,...,p_m$,

$$d(a,g)=\frac{1}{m}\sum_{k=1}^{m}d_{p_k}(a,g). \quad (4)$$

If an article $a$ never occurred in a game with goal $g$ then $d(a,g)$ is undefined. Therefore, our method is incremental, with the number of article associations that are established increasing as more game data is gathered.

An important property of our proposed distance measure is that it is not symmetric: in general, $d(a,b)\neq d(b,a)$ (hence, it is not a distance in the strict geometric sense). Although it could be easily symmetrized (e.g. by taking $\min\{d(a,b),d(b,a)\}$), we do not do this, because asymmetry can be a desirable feature for psychological as well as philosophical reasons [Tversky, 1977]. For instance, $d(\text{MINNEAPOLIS},\text{MINNESOTA})=0.22$, while $d(\text{MINNESOTA},\text{MINNEAPOLIS})=0.12$. Intuitively, this makes sense: when one thinks of MINNEAPOLIS, MINNESOTA is probably one of the first associations, because MINNEAPOLIS is in MINNESOTA. On the flip side, there are many other places in MINNESOTA one could think of, e.g. ST. PAUL, so when thinking of MINNESOTA, MINNEAPOLIS is not as predominant an association. We note that this asymmetry could perhaps be exploited to label concept relationships with their type (e.g. 'is-part-of'). However, we do not address this issue here.

Unlike shortest paths, our measure also does not fulfill the triangle inequality: in general, $d(a,c)\not\leq d(a,b)+d(b,c)$. However, this is actually an asset: Tversky [1977] calls the triangle inequality, which is part of the geometric definition of distance, 'hardly compelling' in the context of semantic distance. Generally, the triangle inequality can be considered to model the transitivity of relatedness. It should be noted that our method can still capture transitive higher-order relatedness, but only when this is suggested by common sense, not by the structure of the graph: even if there is no direct link between two articles, the two will be considered related if people often went through one when aiming for the other.

To conclude the description of our distance measure, we provide an example. Table 1 shows all concepts with a defined distance to NOAM CHOMSKY, in order of increasing distance, i.e. decreasing relatedness. Note that the data comes from only 9 games with goal NOAM CHOMSKY.

## 4 Filtering Unrelated Concepts

Subjectively, Table 1 seems reasonable. The top 6 concepts are all highly related to NOAM CHOMSKY. However, further down the list we have a mix of related and unrelated concepts. One would like to discriminate automatically which of these associations are truly meaningful. Recall from Sec. 2 that the typical anatomy of games is 'get away to hub, then home in on goal'. Since we compute distances to the goal for *all* articles along the path, the articles from the getting-away phase

| | | | |
|---|---|---|---|
| LINGUISTICS | 0.0201 | ~~20TH CENTURY~~ | ~~0.5473~~ |
| COMMUNICATION | 0.0821 | ~~VIETNAM WAR~~ | ~~0.5756~~ |
| LANGUAGE | 0.0896 | ~~ENGLAND~~ | ~~0.6213~~ |
| COMPUTER PROGRAMMING | 0.0985 | ~~UNIVERSITY~~ | ~~0.6620~~ |
| MUSIC | 0.1745 | EDUCATION IN THE U.S. | 0.7401 |
| SOCIALISM | 0.1884 | ~~15TH CENTURY~~ | ~~0.7684~~ |
| ~~SOUND~~ | ~~0.2004~~ | ~~2005 ATL. HURRIC. SEASON~~ | ~~0.8493~~ |
| LIBERAL DEMOCRACY | 0.2155 | UNITED STATES | 0.9431 |
| PHILOSOPHY | 0.2653 | ~~UNITED KINGDOM~~ | ~~0.9598~~ |
| COMPUTER | 0.2747 | ~~HYDE PARK, LONDON~~ | ~~1.0594~~ |
| ENGLISH LANGUAGE | 0.2801 | ~~NORTH AMERICA~~ | ~~1.1376~~ |
| ~~TELEVISION~~ | ~~0.3300~~ | ~~HURRICANE VINCE (2005)~~ | ~~1.1995~~ |
| ~~LA PAZ~~ | ~~0.3465~~ | ~~SPAIN~~ | ~~1.2324~~ |
| COMMUNISM | 0.4130 | ~~EARTH~~ | ~~1.2888~~ |
| ~~ELECTRONIC AMPLIFIER~~ | ~~0.4144~~ | ~~RED DWARF~~ | ~~1.3729~~ |
| ~~RADIO FREQUENCY~~ | ~~0.4195~~ | ~~CANADA~~ | ~~1.4984~~ |
| KARL MARX | 0.4966 | ~~UTRECHT (CITY)~~ | ~~1.6242~~ |

Table 1: Concepts $a$ and $d(a, \text{NOAM CHOMSKY})$. Canceled entries are those eliminated by the method of Sec. 4.

get defined distances to NOAM CHOMSKY, too. However, typically they are no more related to the goal than the hundreds of other concepts whose distances to NOAM CHOMSKY are undefined simply because they never occurred in games with that goal. So, in order to eliminate irrelevant entries, our approach should exclude the articles of the getting-away phase when computing distances.

To get some intuition on how this can be achieved, consider Fig. 2, which shows how the entropies of click probability distributions associated with articles vary along game paths. Since the path length varies among games, we normalized this to $[0,1]$ (the 'normalized goal distance' of the $i$-th article on a path consisting of $n$ articles being $\frac{n-i}{n-1}$). For averaging over all games, we discretized $[0,1]$ into 7 equally sized intervals and computed the means of three quantities for each interval. The left bar is the *prior entropy* $H^0$ of $P^0$. The middle bar is the *posterior entropy* $H^*$ of $P^*$. Entropy measures the uncertainty associated with a distribution, so the right bar $(H^0 - H^*)$ shows the loss of uncertainty afforded by seeing the recorded games. We call this quantity *information gain*.

Games typically follow the pattern 'get away to hub, then home in on goal'. Since all players share the same common sense, they perform these steps in similar ways; e.g., if NOAM CHOMSKY is the goal and a player is currently on LANGUAGE, he/she is much more likely to proceed to LINGUISTICS than to GORILLA. This is why the information gain is high at the start, as players get away to the same hubs, but decreases in the middle of the game; then the gain increases again, as they home in using the same common sense. In other words, the initial getting-away and the final homing-in are much more predictable after seeing game data than before, and the idea is to use the information gain to guess where the homing-in phase, and thus the relevant part of a single game path, starts. An article will then be erased from lists such as Table 1 if it never occurred in the homing-in phase of a game with the given goal. Doing supervised learning, we trained a neural net to predict where the relevant part starts. We comment on its performance next.

# 5 Results

## 5.1 Filtering Unrelated Concepts

Using MTurk [Amazon, 2008], we had human raters mark the *split position* (the article starting the relevant part consisting only of concepts highly related to the goal) of 500 game paths. MTurk is an online platform suitable for such labeling tasks, and it has recently been shown that non-expert labels obtained through it agree very well with gold-standard expert annotations for natural language tasks [Snow *et al.*, 2008]. Every game was labeled by two different people. We used this data to do supervised learning of the most likely split position, using a neural network. Foregoing many details, the network has one hidden layer (two units) and two input features: the number of links between the input article and the article with minimum information gain along the path, and the number of links between the input article and the goal. The class label was 1 if the input article was the one labeled by the human, and 0 otherwise.

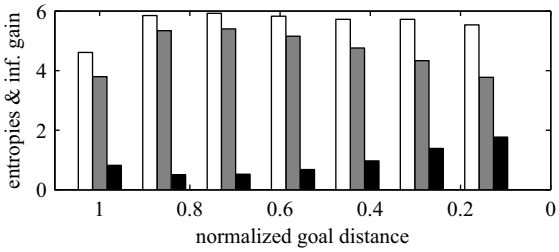


Figure 2: *White:* prior entropy. *Gray:* posterior entropy. *Black:* information gain. (Averages over 1,694 games.)

Once the network is trained, we use it to split unseen paths as follows. For every article along the path, we feed its two features into the network and compute the prediction. We predict the relevant part of the path to start with the article for which the net outputs the highest value. Using cross-validation, we found that our predictor is on average 0.77 positions off the actual split position. This result is good, since the average game consists of as many as 5.7 articles. Moreover, the average inter-human offset was higher, at 0.81. At first glance, the predictor outperforming the humans generating its training data might seem paradoxical; however, this is explained when we look at how often there is an *exact* agreement on the split position: two humans agree for 48% of paths, but human and predictor agree only for 42% of paths. A hand-coded rule that simply splits the path after the article with minimum information gain also did worse, with an average offset of 0.91. The labeled splits and the predictions of the net on the game paths are available online [West, 2009].

## 5.2 Empirical Evaluation of the Distance Measure

In order to test the quality and psychological validity of our distance measure, we compare it to Latent Semantic Analysis (LSA; cf. Sec. 6). We chose LSA because (1) it seems to be the method most widely applied to real-world problems, e.g. automated essay grading [Landauer *et al.*, 1998], (2) it is readily available via a Web interface [Landauer and Kintsch, 1998], and (3) it has been cognitively validated by the psychological community, not only in psychometric but also in behavioral experiments [Huettig *et al.*, 2006].

For this proof of concept [West, 2009], we are using a CD version of Wikipedia [2007] containing around 4,600 articles, but the game could be ported to full-size Wikipedia without a major effort. The articles are stored locally on the game

website and the traces of players during games are stored in a database. In order to gather data that is useful for our purposes, the same goal article is specified in multiple games, played by different players. Initial articles, however, are often chosen at random. We chose $\alpha = 0.1$ in (1).

For this evaluation, the data set contained 1,694 games, collected from players with 282 distinct IP addresses. The set of goals was constrained to 124 randomly selected articles. Each of these 124 target concepts was the goal of between 7 and 26 (median 12) games. For each target, the 5 closest semantic neighbors were picked according to our method and the LSA method, respectively. For LSA, we used the same corpus as Huettig *et al.* [2006]: 'General Reading up to 1st year college' (300 factors). Since we wanted to test for semantic (rather than merely phonetic) relatedness, we did not consider as neighbors words containing the target word or contained in it (e.g. CHOMSKY and NOAM CHOMSKY), the plural of the target, and adjectives directly derived from the target (e.g. CHINA and CHINESE). This yielded usually a set of 10 neighbors for each concept. If both methods agreed on a word, it was included just once, and the neighbor set contained only 9 concepts (this happened for 11 targets). If they agreed on two words, the set contained 8 concepts (this happened for 3 targets). Larger agreements were not encountered. For each target concept, 4 different human raters were given the neighbor set on MTurk (the order of entries in the set was randomized) and asked to select the 3 words they considered most closely related to the target.

Some lists were incorrectly rated (not exactly 3 concepts were selected). Expunging these, 464 rated lists and thus 1,392 selected neighbor concepts remained. Out of these, 64.2% came from our method, while only 32.9% came from LSA and 2.9% of votes went to words suggested by both methods. Clearly, the matches found by our method are preferred by human raters and thus our approach seems to model human common sense better than LSA. The complete results are available online [West, 2009] and summarized in Table 2.

| Method | Votes | Percentage |
|---|---|---|
| Wikispeedia | 893 | 64.2% |
| LSA | 458 | 32.9% |
| Both | 41 | 2.9% |

Table 2: Results of the comparison to LSA.

As a concrete example, consider the concept AIDS: LSA's top 5 neighbors are, in order of increasing distance, ⟨MISCOMMUNICATION*, STALLERS, SPEAKER, LISTENER, NONELECTRONIC⟩. Our method produces ⟨HIV***, WORLD HEALTH ORGANIZATION***, AFRICA**, 20TH CENTURY, INDIA⟩. AIDS was evaluated by 3 raters, and each asterisk stands for one vote. Our method lists exactly the top ranked neighbors first. This example also shows how our method overcomes some of LSA's specific drawbacks. LSA cannot disambiguate between two senses of the same word [Kaur and Hornof, 2005] (SPEAKER appears because the disease cannot be told apart from the plural of AID, the synonym of HELPER), whereas our method is able to differentiate such concepts (Wikipedia article names are already disambiguated). Also, LSA treats every word as representing a single concept, while our method can handle multi-word concepts (Wikipedia article names may contain several words, e.g. WORLD HEALTH ORGANIZATION).

## 6 Related Work

Kaur and Hornof [2005] use existing measures of semantic relatedness (MSRs) to predict user click behavior on websites. Our approach inverts this process by exploiting click behavior in order to construct an MSR. Their paper classifies MSRs into three kinds: taxonomical, statistical and hybrid. Taxonomical measures try to infer relatedness from the structure of a hand-crafted semantic network, most commonly WordNet. Strube and Ponzetto [2006] apply such methods to Wikipedia, restricting their analysis to the tree of category entries and omitting regular articles. On the other hand, statistical MSRs are trained from raw data, e.g. a word association norm or a large corpus. Hybrid approaches combine structural and statistical information. Our method can be described most aptly as hybrid because it is based on Wikipedia's link structure (a very noisy semantic net) but also makes use of human click statistics.

The most widely known corpus-based MSR is LSA [Landauer *et al.*, 1998]. It represents words as points in a high-dimensional vector space constructed from co-occurrence counts in the corpus. Closely related to LSA is Explicit Semantic Analysis [Gabrilovich and Markovitch, 2007]; like our work, it represents concepts as Wikipedia articles, but the overall approach is fundamentally different.

There are other corpus-based MSRs that are not vector-based. Point-wise Mutual Information using Information Retrieval [Turney, 2001] exploits co-occurrence counts in an information-theoretic way, while ICAN [Lemaire and Denhière, 2004] uses them in order to construct an associative semantic network, in which concepts are linked by weighted directed edges. One can think of the MSR computed by our method as representing such a network, too, where concepts with an undefined distance are not linked, and concepts with defined distance are linked by a distance-weighted edge. While there is no apparent intuition for the exact way edge weights change in ICAN, we provide an information-theoretic interpretation for our measure. Our approach also has several properties desirable for cognitive plausibility, as mentioned in the ICAN paper: it is asymmetric, it can account for higher-order relatedness (see discussion in Sec. 3), and it is incremental (new games modify current knowledge).

Recently, Veksler *et al.* [2008] proposed a way of generating a vector space representation from an explicitly defined statistical MSR. Their approach might be applicable in our case, but we have not explored this yet.

Once an MSR has been defined, it can be used for higher-level tasks such as clustering. Wong *et al.* [2007] cluster concepts using primarily Normalized Google Distance (NGD) [Cilibrasi and Vitányi, 2007] and, as a refinement, '*n* degrees of Wikipedia', which is the shortest-path measure described in the introduction. It is interesting to see that, while it is of little value when used stand-alone (as explained), the shortest-path distance can still be helpful in combination with another MSR. NGD is an approximation of the uncomputable Normalized Information Distance (NID). Normally, NID is symmetrized, but the asymmetric definition would be

$K(y|x)/K(y)$, where $K(y)$ is the Kolmogorov complexity of $y$ and $K(y|x)$ the conditional Kolmogorov complexity of $y$ given $x$ [Li and Vitányi, 2008]. Intuitively, this fraction is the percentage of $y$'s information not yet contained in $x$. Kolmogorov complexity is uncomputable, but it can be approximated. NGD makes use of the number of Google hits for the queries "$y$" and "$x,y$" for this purpose. The distance measure we propose can be considered an approximation of the asymmetric NID, too: the numerator in (3) is the number of bits needed to encode a path from $a_i$ to $g$, or in other words, to transform $a_i$ into $g$ (approximating $K(g|a_i)$), while the denominator is the *a priori* number of bits required to encode concept $g$ (approximating $K(g)$).

## 7 Conclusions and Future Work

The main contribution of this paper is a novel method for computing the semantic distance between concepts, based on data from an online game that exploits Wikipedia's hyperlink structure. This approach is inexpensive but cognitively plausible by directly extracting human common sense.

Our measure is computed incrementally; it is asymmetric, accounts for higher-order relatedness, and does not fulfill the triangle inequality, all of which are desirable from a cognitive viewpoint. It also has an information-theoretic interpretation.

We demonstrated the quality of our method by showing that humans rate its performance significantly higher than that of Latent Semantic Analysis in a psychometric evaluation.

Although the incremental character is cognitively plausible, it may be a limitation from the practical viewpoint, in comparison to offline corpus-based methods: we can learn the distance between two concepts only when they co-occur in a game. However, as illustrated, useful concepts are learned even from very few trajectories with a given target. This results in high precision when the task is to find semantic neighbors of concepts that occurred as goals, as shown experimentally; due to the sparse coverage, however, recall would presumably be smaller if the task were to compute the relatedness of two arbitrary concepts that might not have co-occurred in any game. We are currently addressing this problem by investigating dimensionality reduction techniques for generalizing the distance measure to unseen concept pairs.

In future work, we plan to investigate the use of games not only to weight concept relationships, but also to determine their type, a further step towards automatically extracting a rich semantic network from freely available Web data.


### Acknowledgements

This research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).


## References


[Amazon, 2008] Amazon. Amazon Mechanical Turk. Website, 2008. http://www.mturk.com.

[Brin and Page, 1998] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Computer Networks and ISDN Systems*, 1998.

[Cilibrasi and Vitányi, 2007] R. L. Cilibrasi and P. Vitányi. The Google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3), 2007.

[Dolan, undated] S. Dolan. Six degrees of Wikipedia. Website, undated. http://www.netsoc.tcd.ie/~mu/wiki (accessed Dec. 23, 2008).

[Fellbaum, 1998] C. Fellbaum, editor. *WordNet. An Electronic Lexical Database*. MIT Press, 1998.

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, 2007.

[Huettig *et al.*, 2006] F. Huettig, P. T. Quinlan, S. A. McDonald, and G. T. M. Altmann. Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121(1), 2006.

[Jiang and Conrath, 1997] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*, 1997.

[Kaur and Hornof, 2005] I. Kaur and A. J. Hornof. A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. In *CHI*, 2005.

[Landauer and Kintsch, 1998] T. Landauer and W. Kintsch. LSA. Website, 1998. http://lsa.colorado.edu.

[Landauer *et al.*, 1998] T. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 1998.

[Lemaire and Denhière, 2004] B. Lemaire and G. Denhière. Incremental construction of an associative network from a corpus. In *CogSci*, 2004.

[Li and Vitányi, 2008] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 3rd edition, 2008.

[Rada *et al.*, 1989] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.*, 19(1), 1989.

[Snow *et al.*, 2008] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP*, 2008.

[Strube and Ponzetto, 2006] M. Strube and S. P. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI*, 2006.

[Turney, 2001] P. D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML*, 2001.

[Tversky, 1977] A. Tversky. Features of similarity. *Psychological Review*, 84(2), 1977.

[Veksler *et al.*, 2008] V. D. Veksler, R. Z. Govostes, and W. D. Gray. Defining the dimensions of the human semantic space. In *CogSci*, 2008.

[von Ahn *et al.*, 2006] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *CHI*, 2006.

[West, 2009] R. West. Wikispeedia. Website, 2009. http://www.wikispeedia.net.

[Wikipedia, 2007] Wikipedia. 2007 Wikipedia Selection for schools. Website, 2007. http://schools-wikipedia.org (accessed Aug. 3, 2008).

[Wikipedia, 2009] Wikipedia. Wiki Game. Website, 2009. http://en.wikipedia.org/w/index.php?title=Wikipedia:Wiki_Game&oldid=281%582697.

[Wong *et al.*, 2007] W. Wong, W. Liu, and M. Bennamoun. Tree-traversing ant algorithm for term clustering based on featureless similarities. *Data Min. Knowl. Discov.*, 14(3), 2007.