Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia

Philipp Singer[a], Thomas Niebler[b], Markus Strohmaier[c,d], Andreas Hotho[b]

[a] *Knowledge Technologies Institute, Graz Technical University, Graz, Austria*
[b] *Data Mining and Information Retrieval Group, University of Würzburg, Würzburg, Germany*
[c] *Computer Science Institute, University of Koblenz-Landau, Koblenz, Germany*
[d] *Computational Social Science Group, GESIS, Cologne, Germany*

Abstract

In this article, we present a novel approach for computing semantic relatedness and conduct a large-scale study of it on Wikipedia. Unlike existing semantic analysis methods that utilize Wikipedia's content or link structure, we propose to use *human navigational paths* on Wikipedia for this task. We obtain 1.8 million human navigational paths from a semi-controlled navigation experiment – a Wikipedia-based navigation game, in which users are required to find short paths between two articles in a given Wikipedia article network. Our results are intriguing: They suggest that (i) semantic relatedness computed from human navigational paths may be more precise than semantic relatedness computed from Wikipedia's plain link structure alone and (ii) that not all navigational paths are equally useful. Intelligent selection based on path characteristics can improve accuracy. Our work makes an argument for expanding the existing arsenal of data sources for calculating semantic relatedness and to consider the utility of human navigational paths for this task.

Keywords:
semantic relatedness, navigation, Wikipedia

Email addresses: `philipp.singer@tugraz.at` (Philipp Singer), `thomas.niebler@uni-wuerzburg.de` (Thomas Niebler), `markus.strohmaier@gesis.org` (Markus Strohmaier), `hotho@informatik.uni-wuerzburg.de` (Andreas Hotho)

This article appears in International Journal of Semantic Web Information Systems, vol 9(4), edited/authored by Amit Sheth, Copyright 2013, IGI Global, http://www.igi-global.com/journal/international-journal-semantic-web-information/1092. Reprinted by permission of the publisher.

1. Introduction

Computing *semantic relatedness*[2] between concepts represents a fundamental challenge on our way to a semantically-enabled web. Especially, common sense knowledge in terms of semantic relatedness is of special interest in e.g., improving information retrieval or language processing. To obtain a judgement of semantic relatedness of two terms or concepts, the idea is to rely on the accumulated or common knowledge. Rubenstein & Goodenough (1965) have pointed out that there is a positive relationship between the degree of semantic relatedness of a pair of terms and the degree to which their contexts are similar. Hence, the idea is that a semantic relatedness score captures this common sense knowledge over a set of contexts and abstracts and generalizes it.

Psychological experiments (Tversky, 1977; Medin et al., 1993) have shown that semantic relatedness is both *context dependent* and *asymmetric*. Context dependency means that the determined relatedness is influenced by the context the words appear in and the semantic relatedness may be asymmetric as people may provide distinct ratings depending on the direction the words are presented. Nevertheless, Aguilar & Medin (1999) showed that this asymmetry just occurs at special occasions and Medin et al. (1993) also showed that the difference in ratings for a given word pair is less than five percent. Hence, we will focus on *symmetric semantic relatedness* in this work, as we believe that this is sufficient for the investigations we want to conduct and we can ignore these small differences.

Recent approaches to identify semantic associations between concepts exploit the rich fabric of emerging information networks such as Wikipedia. Existing semantic analysis methods such as those by Gabrilovich & Markovitch (2007), Ponzetto & Strube (2007a) or Yeh et al. (2009) have shown great potential by using textual or structural (link) information on Wikipedia. While these methods have produced promising results, they only capture semantics from a limited set of people (e.g., Wikipedia editors) and they mostly neglect pragmatics (i.e., how Wikipedia is used). At the same time, millions of web users navigate Wikipedia daily to find information, to educate themselves or for research issues. When navigating a set of articles on Wikipedia, users typically need to tap into their intuitions about real-world concepts and the perceived relationships between them in order to progress towards their set of targeted articles. Humans tend to find intuitive paths instead of necessarily short paths, while contrary an automatic al-

[2] Note that semantic *relatedness* does not necessarily mean the same as *similarity*. Amongst others it includes: similarity, meronymy, hypernymy or IS-A relationships.

gorithm would try to find a shortest path between two concepts that may not be as semantically rich and intuitive as a navigational path conducted by a human.

A great advantage of such navigational paths by humans is that they can be captured in a very simple way. The only prerequisite is that there is a group of users that navigate a system. Furthermore, many existing methods only work well if the system at hand provides high quality content that can be leveraged for calculating semantic relatedness. Contrary, our approach is independent of the content of a resource. It also gives opportunities to calculate semantic relatedness between different kind of resources. For example, suppose we want to calculate semantic relatedness between images and textual pages of a website. This would be a very difficult task for content based approaches, as both resources exhibit different features. The method proposed in this work though would work on any type of resource as long as it is navigated by users.

While such data about navigational paths could potentially represent a profoundly rich resource for calculating semantic relatedness between concepts, it has not received much attention by the research community yet.

1.1. Research Questions

Consequently, we would like to explore (i) whether human navigational paths represent a useful resource for calculating semantic relatedness between concepts on Wikipedia at all, and (ii) if so, in what ways, e.g., what kinds of navigational paths are particularly useful?

In this paper, we tackle these questions and present a series of principled experiments studying the usefulness of almost 1.8 million human navigational paths on Wikipedia for calculating semantic relatedness between concepts (cf. our previous work on this topic (Singer et al., 2013)). Navigational data was obtained from a semi-controlled, large-scale navigation experiment – a Wikipedia-based game called "The WikiGame"[3], in which users need to navigate from a given Wikipedia concept (the starting node) to another concept (the target node). These human navigational Wikigame paths present an abstraction of real user navigation in information networks and enable us to give detailed insights into the usefulness of such data[4].

[3] http://www.thewikigame.com

[4] When we speak about human navigational paths throughout or experiments we refer to the paths captured via the game.

1.2. Contributions of the paper

Our experiments demonstrate that human navigational paths – captured via a Wikipedia-based navigation game – can represent a viable source for calculating semantic relatedness between concepts in information networks. We show that semantic relatedness calculated on this kind of human navigational data can be more precise than semantic relatedness calculated on paths automatically extracted from Wikipedia's plain link structure. Finally, we find that not all navigational paths are equally useful. Intelligent selection of navigational paths based on path characteristics can improve accuracy.

The paper is structured as follows: In Section 2, we give an overview of related work. Section 3 describes our methodology for calculating semantic relatedness based on navigational paths together with a description of the datasets and evaluation methods that we have used in this work. This is followed by Section 4, where we conduct baseline experiments to explore whether human navigational paths can contribute to the task of computing semantic relatedness. In Section 5, we present results from path selection experiments where we investigate which characteristics of human navigation paths render them useful for semantic relatedness. Finally, we discuss and conclude our work in Section 6.

2. Related Work

Computing semantic relatedness between concepts has received much attention from our research community in the last few years, and a wide array of approaches exists. Semantic relatedness scores are widely needed and used in a variety of applications and studies, e.g., word sense disambiguation (Resnik, 1998), usage for word spelling errors (Budanitsky & Hirst, 2001), text segmentation using lexical cohesion (Kozima, 1993; Manabu & Takeo, 1994), image (Smeulders et al., 2000) or document (Srihari et al., 2000) retrieval, cognitive science (Talmi & Moscovitch, 2004) and many more. For a great overview over many different methods to calculate semantic relatedness, see the survey done by Zhang et al. (2012).

Li et al. (2003) point out that semantic relatedness measures and methods can basically be categorized into two groups: *edge-counting-based* and *information-theory-based* methods. When we suppose that a lexical taxonomy has a tree shape then Rada et al. (1989) proved that the distance in the minimum number of edges that separate two given words in such a tree is a metric for specifying the semantic distance between these two words – or to be more precise: the semantic relatedness. While these edge-counting methods make use of *IS-A* relations only, they are very useful for applications with highly constrained taxonomies (Li et al., 2003). According to Resnik (1998) the information-theory-based methods define semantic relatedness between two words using information content and the more information two concepts or words share the more related they are. Jiang & Conrath (1997) presented an approach for computing semantic relatedness between words and concepts combining both edge-based and information-theory-based methods. This method is often simply referred to as the *Jiang-Conrath distance*.

Above described methods can be applied to different information resources. One of the most often and successfully used resource for calculating semantic relatedness is the lexical database WordNet[5] (Miller, 1995). Yang & Powers (2005) proposed a new methodology for calculating semantic relatedness on WordNet using edge-counting techniques. In (Patwardhan, 2006) the authors introduced a WordNet based measure of semantic relatedness by combining both structure and content of WordNet and furthermore enhanced it with co-occurrence information derived from raw text. This enabled the authors to build *gloss* vectors and hence, they used cosine similarity in order to specify semantic relatedness scores between words. A similar approach has been conducted by Banerjee & Pedersen (2003) who used glosses to determine the number of shared words between the definitions of two words for specifying the semantic relatedness between them. Budanitsky & Hirst (2001) compared five different measures of semantic relatedness on WordNet and concluded that the Jiang-Conrath distance is the most accurate by evaluating the results against human judgements and an actual NLP task. Pedersen et al. (2004) introduced a PERL module that implemented nine different measures of semantic relatedness using WordNet and it is widely used by researchers. In subsequent work by Budanitsky & Hirst (2001) the authors again evaluated several semantic relatedness measures using the introduced PERL module using the task of detecting and correcting real-world spelling errors. The authors again show that the Jiang-Conrath distance is superior to other methods. Navigli & Ponzetto (2012a) took WordNet one step further by creating BabelNet, an automatically generated multilingual extension of WordNet. In their publication, they covered the generation of BabelNet by incorporating WordNet, Wikipedia and Machine Translation tools, its evaluation on both new and existing gold standard datasets and the viability to use BabelNet as a resource to perform both monolingual and cross-lingual word sense disambiguation (see Navigli & Ponzetto (2012b)).

More recently, with the rise of the Web 2.0, user-generated content provided

[5] http://wordnet.princeton.edu/

great opportunities for calculating semantic relatedness scores by directly leveraging data generated by humans. Especially tagging systems have attracted lots of interest as a source of data for this task in the past (e.g., (Strohmaier et al., 2012), (Helic et al., 2011), (Cattuto et al., 2008) or (Markines et al., 2009)). But also information networks like Wikipedia have received attention as a resource for calculating semantic relatedness. Because giving a complete review of the literature in this vast field of calculating semantic relatedness using user generated content is beyond the scope of this paper, we will primarily focus our discussion on a few algorithms and methods that are most salient and relevant to this work. Instead, we point the interested reader to a capacious survey about the uses of Wikipedia for many purposes done by Hovy et al. (2012).

Many of the methods we discuss here have been developed for or can easily be applied to Wikipedia. In the following, we differentiate between methods which focus on exploiting different aspects of information networks such as Wikipedia – especially *content* and *links*.

2.1. Content-based methods

A simple way of determining the relatedness between concepts is to represent the content of Wikipedia articles as bag-of-word vectors (Manning et al., 2008). Relatedness between two concepts can then be computed by calculating the similarity between vectors by e.g., using *cosine similarity*.

Gabrilovich & Markovitch (2007) applied *tf-idf* to Wikipedia and introduced a method called *Explicit Semantic Analysis (ESA)*. This method builds a weighted inverted index and extracts a weighted vector of Wikipedia concepts. The vectors of different concepts can be compared, which leads to a calculation of relatedness between terms based on their *tf-idf* weighted vectors. One of the advantages of ESA is that it allows to calculate the relatedness between arbitrary text – e.g., individual words or long documents.

Another method for calculating semantic relatedness is *Latent Semantic Analysis (LSA)* (Landauer et al., 1998; Deerwester et al., 1990). LSA can be used for determining semantic relatedness between Wikipedia concepts by producing word count matrices based on articles and reducing their dimensionality using *singular value decomposition*. Similarity again can be calculated using the angle between vectors.

In addition to analyzing content, link based methods have received increasing attention by our research community lately.

2.2. Link-based methods

Two main types of link based methods can be distinguished: (a) methods focusing on link information present for a specific page – i.e., links on a page can be seen as some type of topical markers – and (b) methods exploiting paths through Wikipedia's link network.

2.2.1. Links as topical markers for Wikipedia concepts

Ito et al. (2008) use co-occurrence information between links present on the same page for computing semantic relatedness between concepts using a co-occurrence window size of *k* and pruning the vectors with a *tf-idf* based approach. Milne (2008) has proposed a new method of calculating semantic relatedness on Wikipedia leveraging the link structure called "The Wikipedia Link Vector Model". This model judges the similarity between two articles by calculating the angle between the link vectors between two pages. The vectors are built by link counts weighted by the probability of each link occurring. Furthermore, the links get an additional weighting to reduce the impact of frequently occurring links to very common target concepts. Turdakov & Velikhov (2008) have established a similar approach to exploit Wikipedia's link structure in order to calculate similar Wikipedia pages. The technique uses Dice's measure and also ranks two pages similar, if the fraction of similar links is high. The authors as well use a different weighting scheme for the type of link that occurs on a page and they evaluate their approach based on a *word-sense disambiguation* task showing that they can achieve better results than a naive technique of just looking at the neighborhoods of the context and the term in Wikipedia. A more recent method is *Salient Semantic Analysis (SSA)* (Hassan & Mihalcea, 2011). SSA leverages salient features in the context of a term. For example, links on Wikipedia can be interpreted as salient features for terms inside some predefined distance.

2.2.2. Topology based methods

Ito et al. (2008) have introduced an adaption to tf-idf called *pfibf* utilizing links between two concepts inside Wikipedia's link network. The assumption is that (i) the number of paths from article *i* to *j* in the Wikipedia topology and (ii) the length of each path from article *i* to *j* determine the relatedness between two concepts.

In (Yeh et al., 2009) the authors present *WikiWalk*, a method that performs random walks based on Personalized PageRank. Based on the output vectors of individual random walks for given words, semantic relatedness is calculated by computing the similarity between both vectors. By pruning the initialization of the teleport vector with Explicit Semantic Analysis, the authors report that their method can even slightly outperform ESA.

Yazdani & Popescu-Belis (2013) created a network topology by parsing the contents of Wikipedia articles and linking articles which are semantically similar. They applied a weighted random walk technique on both the artificially created network as well as the basic Wikipedia topology and calculated the *visiting probability* from one set of nodes to another. They finally showed that a combination of both techniques performed better than both techniques alone.

Strube & Ponzetto (2006) show that straightforward path based measures work very well when focusing on Wikipedia's category taxonomy and that a combination with WordNet is very suitable in order to improve the corresponding accuracy. Furthermore, the authors have evaluated their results by performing a NLP based case study, showing that such knowledge bases collaboratively produced by a huge amount of users like Wikipedia actually can be used for such tasks with similar effects to hand-crafted taxonomies by experts like WordNet (see also (Ponzetto & Strube, 2007b)). In (Milne & Witten, 2008) the authors proposed a similar approach called the "Wikipedia Link-based Measure (WLM)" which as well only leverages Wikipedia's hyperlink structure while it ignores the content and category hierarchy. In (Ponzetto & Strube, 2007a) the authors extend their idea by automatically determining *isa* and *notisa* relations between Wikipedia categories. An automatic extraction of the type of semantic relations has also been successfully conducted by Nakayama et al. (2008).

The work most related to this paper is by West et al. (2009), who have analyzed a set of human navigational paths obtained from *Wikispeedia*[6], a game similar to *"TheWikiGame"*. The authors introduce a method for computing an asymmetric relatedness measure for concepts based on human navigational paths in the corpus. The authors focus on calculating semantic relatedness based on information between a concept in a path and the target page of this game. To the best of our knowledge, West et al. (2009) have been the first to study semantics in human navigational paths on Wikipedia. While their work demonstrates the great potential of this approach, it is limited in some ways: (i) semantic relatedness can only be calculated between a node in a path and a specific target node of a game if they directly co-occur in a path or (ii) the dataset was limited to a small subset of Wikipedia and to a comparatively small set of navigational paths – concretely 1,694 paths.

[6] http://www.cs.mcgill.ca/ rwest/wikispeedia/

2.3. Summary

Calculating semantic relatedness has proven to be an important facet needed for several applications. Many researchers focused on leveraging lexical taxonomies for calculating semantic relatedness scores. More recently, our research community also proposed methods for using user generated content like tagging data or information networks like Wikipedia. As many existing state-of-the-art works evaluated their methods on the same WordSimilarity-353 gold standard dataset, we report some previous accuracy results in Table 1. However, we believe that it is difficult to directly compare our accuracy results to those obtained by existing well-known methods as the exact evaluation mechanisms of existing methods are difficult to judge. We provide a short discussion about this topic in Section 6.

Recent research on link and path based measures (e.g., (Ito et al., 2008), (Yeh et al., 2009) or (Strube & Ponzetto, 2006)) has demonstrated the potential of exploiting topological link structure of Wikipedia for determining semantic relatedness. Our work significantly expands the state-of-the-art in this area by presenting a method for calculating semantic relatedness that utilizes data about human navigational paths through Wikipedia's topological link network. We build on the work and first signals detected by West et al. (2009), but use a novel approach for calculating semantic relatedness based on a corpus of navigational paths that overcomes several limitations the method of West et al. (2009) exhibits. Concretely, the method conducted in this paper can calculate semantic relatedness between any two nodes in a corpus of paths and not only between a node in a path and a specific target game node. We also overcome the necessity of a direct co-occurrence in at least one path between two nodes if one wants to determine the semantic distance between these two concepts. The only limitation of our methodology is that a concept is present at least once in any single path of the corpus in

Table 1: WordSimilarity 353 scores for existing methods

Method	Score	Reference
WikiRelate!	0.48	(Strube & Ponzetto, 2006)
LSA	0.56	(Finkelstein et al., 2002)
WikiWalk	0.63	(Yeh et al., 2009)
WordNet	0.66	(Agirre et al., 2009)
WLVM	0.72	(Milne, 2008)
ESA	0.75	(Gabrilovich & Markovitch, 2007)

order to calculate the semantic relatedness between this and any other concept. In particular, we i) expand the scope of current investigations dramatically (we use ∼*1.8 million paths* from games that are taking place on the *entire* English Wikipedia), ii) deploy state-of-the-art evaluation techniques (WordSimilarity-353 and other standard evaluation datasets) and iii) identify characteristics of navigational paths that are most useful for computing semantic relatedness.

3. Methods and Datasets

In the following, we establish some preliminaries for our work; then we discuss different relatedness measures and the way we apply them to our corpus of human navigational paths. Finally, we describe the datasets at hand and our evaluation method.

3.1. Preliminaries

We define a Wikipedia $\mathbb{W}$ graph G as a graph $G_{\mathbb{W}} = (V_{\mathbb{W}}, E_{\mathbb{W}})$ with vertices – i.e., pages or concepts – $V_{\mathbb{W}}$ and directed edges – i.e., links – $E_{\mathbb{W}} = \{(v, w)|v, w \in V_{\mathbb{W}}\}$. A page $v = (id, title, content) \in V_{\mathbb{W}}$ is a triple of a positive integer *id*, denoting an unique number for easy identification, a string *title*, denoting the title of the page (name) as well as another string *content*, which contains a definition as well as a description of the concept given by the title. The content also contains all the links which define the edges originating from this page. In fact, an edge (v, w) can only be contained in $E_{\mathbb{W}}$, iff the *content* of page v contains a hyperlink to page w.

We can now define $inlinks(v)$ and $outlinks(v)$ for a given page v. The set of outlinks contains all links originating from v and is easily deduced as $outlinks(v) = \{(v, w) \in E_{\mathbb{W}}|w \in V_{\mathbb{W}}\}$. The set of inlinks contains all links pointing from different pages to page v and is defined analogously as $inlinks(v) = \{(w, v) \in E_{\mathbb{W}}|w \in V_{\mathbb{W}}\}$, but is not as directly tractable as the set of outlinks.

Given a graph $G = (V, E)$ (e.g., a Wikipedia graph $G_{\mathbb{W}}$) with vertices V and directed edges $E = \{(v, w)|v, w \in V\}$, we now define a *path* p as a n-tuple $(v_1, \ldots, v_n)$ with $v_i \in V, 1 \leqslant i \leqslant n$ and $(v_i, v_{i+1}) \in E, 1 \leqslant i \leqslant n - 1$. We define $\mathbb{P}$ as the set of all paths and the length of a path $len(p)$ as the length of the corresponding tuple $(v_1, \ldots, v_n)$. Additionally, we want to define $\mathbf{p} = \{v_k|k = 1 \ldots n\}$ as the set of nodes in a path p. Note that $|\mathbf{p}| \leqslant n$.

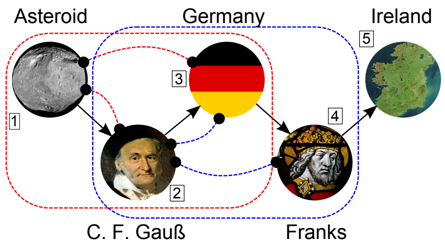
3.2. Measures for semantic relatedness

Schuetze & Pedersen (1997) introduced the method for calculating semantic similarity using lexical co-occurrence information between words – or in our case Wikipedia concepts. The basic idea is to represent each concept as a vector capturing the co-occurrence count to all other concepts in a multi-dimensional space.

A simple procedure for determining semantic relatedness between concepts based on such co-occurrence information is to use *direct co-occurrence*. *First-order co-occurrence* (Schuetze & Pedersen, 1997) implies that concepts can only be similar if they co-occur directly (e.g., in the same documents or in our case paths). However, in our experiments we have observed that this way of calculating semantic relatedness is not suitable for navigational data because many highly related concepts never directly appear in the same path. Furthermore, many word pairs of the WordSimilarity-353 evaluation dataset never co-occur directly in our available data. Also, first-order co-occurrence focuses on semantic relatedness with a tendency to more general concepts.

To avoid this problem, we calculate relatedness between concepts based on the similarity between their corresponding co-occurrence vectors. This is referred to as *second-order co-occurrence* (Schuetze & Pedersen, 1997), which assumes that words are semantically related if they share similar neighbors. Second-order co-occurrence emphasizes if two concepts i and j are similar in a synonymous way. This method also removes the necessity of two concepts directly co-occurring in a path for specifying the semantic relatedness between them and is one of the main advantages of our method over the one introduced in (West et al., 2009). We will use this method for the purpose of our paper.

In order to be able to calculate second-order co-occurrence similarity between two Wikipedia concepts i and j, the corresponding vectors $v_i = [co_{i1}, co_{i2}, ..., co_{in}]$ and $v_j = [co_{j1}, co_{j2}, ..., co_{jn}]$ for both concepts are required. In both vectors, co_{ik} or co_{jk} is the corresponding first-order co-occurrence count between concepts i and k or j and k. We can calculate the relatedness between vectors v_i and v_j by using a similarity (distance) measure between vectors. As an example, let us suppose we want to calculate the semantic relatedness between concept $i = Germany$ and $j = Ireland$ given our example illustrated in Figure 1a. We use the corresponding vectors v_i and v_j present in the symmetric co-occurrence matrix v depicted on the right side in Figure 1b and calculate the cosine similarity measure given both vectors, which results in 0.35 for this simple example (the sliding windows mechanism will be described in Section 3.3). Throughout this work we use *cosine similarity* (Cattuto et al., 2008; Salton, 1989) which has linear complexity and has shown good performance in comparable cases. The choice of similarity measures is secondary to our method[7].



(a) Sliding windows

	Asteroid	C. F. Gauß	Germany	Franks	Ireland
Asteroid	0	1	1	0	0
C. F. Gauß	1	0	1	1	0
Germany	1	1	0	1	1
Franks	0	1	1	0	1
Ireland	0	0	1	1	0

(b) Co-occurrence matrix

Figure 1: Figure 1a illustrates the sliding window mechanism for a window size of $k = 3$ on a path from *Asteroid* to *Ireland*[8]. Circles represent Wikipedia articles, rounded rectangles represent a window. The solid arrows represent the path taken, the dashed lines with dotted ends each represent a (symmetric) co-occurrence between two concepts. We only highlight the first two windows. The resulting co-occurrence matrix after all steps is shown on the right in Figure 1b.

3.3. Semantic relatedness for paths

To compute semantic relatedness using co-occurrence information inside a corpus of navigational paths, we define a *co-occurence graph* between concepts as a *weighted undirected graph* $G_{coocc} = (V_{\mathbb{W}}, E_{coocc})$ where the set of vertices $V_{\mathbb{W}}$ are the corresponding Wikipedia concepts available for all paths in the corpus. The set of edges E_{coocc} is defined as follows: An edge $e = \{u, v\}$ lies in E_{coocc}, iff u and v appear on the same path p, i.e., if $u, v \in p$. The weight of the edge $w(e)$ is determined by the number of co-occurrences of u and v on any path $p \in \mathbb{P}$. We use undirected co-occurrence edges as we do not want to explicitly capture the order of the appearance of two nodes in a path but rather specify their symmetric co-occurrence as we are also calculating symmetric semantic relatedness between

[7] We used other vector similarity measures like *Mutual Information* or *Dice Coefficient* with similar results.

[8] The asteroid picture is courtesy of NASA/JPL-Caltech. All other pictures are published under the Creative Commons licence.

two concepts. To capture relatedness of two concepts in a corpus of human navigation paths, we use *sliding windows* of a variable size k following the natural assumption that the distance between two concepts is crucial for calculating precise semantic relatedness scores (cf. (Schuetze & Pedersen, 1997)). In this paper we investigate paths instead of documents. Hence, we follow and investigate the hypothesis that the navigational distance between two concepts in a path (i.e., they are just a specified hop range away in a path) is important in order to calculate precise semantic relatedness scores. Given a navigational path with a large length of 20 visited nodes, it may make more sense to consider the co-occurrence between the first and third node in the path instead of the first and final target node in this long path.

Formally, this sliding window process can be expressed in the following way: An undirected co-occurence edge $e = \{u, v\}$ between two concepts $u, v \in V_{\mathbb{W}}$ only exists, iff u and v appear on the same path p and for the directed subgraph $q = (u, \ldots, v)$ of p the inequality $len(q) \leqslant k$ holds. Figure 1a illustrates how we calculate the co-occurrence between concepts available in a path with a sample window size of $k = 3$. The red box represents the first window of the path (left-most window) in which the concept *Asteroid* co-occurs with the next $k - 1 = 2$ concepts in the path (*C.F. Gauss* and *Germany*). Since we use a symmetric co-occurrence measure, the next two concepts co-occur with *Asteroid* as well. The window then slides one step to the right, (blue box, right most window). We repeat this step until position n is reached. The resulting co-occurrence matrix is shown in Figure 1b – higher co-occurence counts are possible for larger data. Using this matrix, the relatedness between concepts can be determined by calculating a similarity (i.e., cosine) between two concept-vectors (see Section 3.2).

3.4. Datasets

We now introduce the datasets for our experiments and the ways in which they have been obtained.

3.4.1. Wikigame

This dataset is based on the online game *"TheWikiGame"*[9]. The platform offers users a multiplayer game, where the goal is to navigate from one Wikipedia page (the start page) to another Wikipedia page (the target page) which is linked to the start page through Wikipedia's underlying topological link network. The users

[9] http://thewikigame.com/

can leverage Wikipedia's directed link structure to reach their target node, but in some cases users also establish new links in their paths between articles that might not yet exist in Wikipedia's topological link network. This can happen when, for example, users use the back button in their browser to navigate to a previous article, and the current article does not have a link back to the previous one. One explanation for such behavior could be that users originally end up at a concept they are not happy with and decide that going another route may be a better idea. This is a rich feature of this dataset as it enables us to establish relations between concepts that we normally would not see using Wikipedia's link network. The logic of the game can be transported to any information network consisting of links between resources. If the user is presented with all links leading from one page to another the game can be applied and played in similar fashion.

A path in this dataset is the attempt of a single player to solve a game. We only consider paths where a user navigates through at least two pages and only if those pages are available in our Wikipedia dump consisting of concepts from the *main namespace* (see Section 3.4.2). Furthermore, we know which paths are *successful* – i.e., the user has reached the target concept – and which are *unsuccessful* – i.e., the user has failed to find a route to the target in the given timeframe. Table 2 shows some main characteristics of our Wikigame dataset. The adjusted dataset at hand consists of 1,799,015 navigation paths captured between 2009-02-17 and 2011-09-12. The distribution of path lengths is discussed and depicted later in Figure 3. We can see differences in the length distribution for successful, unsuccessful paths and all paths, but each distribution exhibits a peak at a length of around six.

3.4.2. Wikipedia

Wikipedia offers complete dumps of the English Wikipedia, and for our experiments, we chose a Wikipedia dump dated on 2011-11-07. The reason for this choice was that this was the dump closest to the timestamps in the Wikigame dataset (see Section 3.4.1) that was publicly available[10]. We obtained the present page-to-page network provided by this dump and limited it to links between pages from the main namespace and also to links between the distinct pages available in our Wikigame dataset. The reason for this is that we want to compare the paths through the network with the corresponding topological network; if we would leave the original network untouched, it would be impossible to assert whether

[10] Wikipedia only makes a specific amount of recent dumps available for download

the difference in our results are based on the type of paths or on the number of distinct pages in the Wikigame dataset.

3.5. Evaluation

To evaluate semantic relatedness, we compare our results to a gold standard dataset, specifically the *WordSimilarity-353* dataset (Finkelstein et al., 2002). The WordSimilarity-353 dataset consists of 353 pairs of English words and names and includes all 30 nouns of the *Miller and Charles dataset* (Miller & Charles, 1991) and most of the 65 pairs of the *Rubenstein and Goodenough dataset* (Rubenstein & Goodenough, 1965). Each pair was assigned a relatedness value between 0.0 (no relation) and 10.0 (identical), denoting the assumed common sense semantic relatedness between two words. For each pair of words, ratings of 16 different people were collected. Finally, the total rating per pair was calculated as the mean value of each of the 16 user's ratings. This way, WordSimilarity-353 provides a valuable evaluation base for comparing our concept relatedness scores computed on Wikipedia to an established human generated and validated collection of word pairs. In (Miller & Charles, 1991) it was also shown that the correlation coefficient between the two sets of ratings – i.e., the *Miller and Charles dataset* and the *Rubenstein and Goodenough dataset* – is 0.97. Hence, we can conclude that human knowledge about semantic similarity between words is very stable over a large time span and we can use them for evaluating our semantic relatedness

Table 2: Characteristics of TheWikiGame dataset

#Pages	360,417
#Games	361,115
#Users	260,095
#Paths	1,799,015
#Visited nodes	10,758,242
Average path length	5.98
Average #paths per user	6.92
#Successful paths	653,081
#Visited nodes of successful paths	4,116,879
Average successful path length	6.30
#Unsuccessful paths	1,145,934
#Visited nodes of unsuccessful paths	6,641,363
Average unsuccessful path length	5.80

calculations (Li et al., 2003).

Because WordSimilarity-353 consists of English words and names, we map them to an according Wikipedia concept. We use an adapted version of WordSimilarity-353 called *WikipediaSimilarity-353*, which contains a manual mapping and disambiguation step of words contained in WordSimilarity-353 to Wikipedia concepts (Milne & Witten, 2008). As a further step, we manually checked the mappings for correctness and modified some of the mappings accordingly[11]. For some word pairs it is not possible to map it to appropriate Wikipedia concepts[12]. By removing such pairs where we can not map one word, we end up with 314 concept pairs where we can cover a total of 308 pairs with the concepts available in our Wikigame and Wikipedia dataset (see Sections 3.4.1 and 3.4.2). The main reason for our choice of using manual mappings instead of for example, using sense pairs with maximal similarity, is that the main focus of our work is to show the viability of human navigational paths for calculating semantic relatedness and not the necessarily best working method to date. Milne (2008) shows in his work that the accuracy drops by a large margin if one does not use a manual mapping and relies on automatic disambiguation. This automatic disambiguation step itself is not trivial and can probably introduce a large negative bias to our results as this would make inference of the results difficult as we would not know if the possibly bad results are based on the simple disambiguation step or on the bad results of our method. In the remaining chapters, we will refer to *WordSimilarity-353* even if technically, we mean *WikipediaSimilarity-353*. Our final mapping can be found online on our website[13].

Finally, we compare two rankings. We extract the first ranking of the original scores available through WordSimilarity-353. We also create a similarity ranking for the corresponding word pairs on different paths corpora with our semantic relatedness method, using the cosine similarity. In the last step, we compare both rankings with the Spearman rank correlation coefficient as stated in Formula 1. Using the Spearman rank correlation as evaluation metric enables us to specify how closely our semantic relatedness scores are in terms of a ranked list on all WordSimilarity-353 concept pairs. If the rank correlation is close to 1 we nearly produce the same ranking as human judges.[14]

[11] For example, we had to correct some Wikipedia ids of concepts.

[12] For example there are no appropriate Wikipedia pages available for both the terms in the word pairs “Hotel reservation” or “Boxing round”.

[13] http://www.philippsinger.info/wikisempaths.html

[14] One needs to note that the smaller the gold standard is one compares to, the more difficult it

$$\rho = \frac{Cov(rg_{WS}, rg_{WP})}{\sigma_{rg_{WS}} \sigma_{rg_{WP}}} \in [-1;1] \tag{1}$$

In this formula, rg_{WS} refers to the ranks in WordSimilarity-353, and rg_{WP} to our results. The σ_{rg_X} values is the standard deviation of both ranks. Bear in mind that ranks can also contain tied values, i.e., where two word pairs share the same similarity value. We made sure that our implementation can also handle such ties. We also calculated significance using a two-sided p-value which roughly indicates the probability that a uncorrelated system produces a ranking that has at least the same Spearman rank correlation as the one computed from the original ranking produced by our method. We will not explicitly specify the p-values for each calculation, as all p-values are below the significance level of 0.01. Hence, when we talk about the Spearman rank correlation, we actually refer to the calculated ρ.

4. Semantics of navigational paths

To study feasibility, we first investigate whether a corpus of human navigational paths through an information network – i.e., navigational paths taken from the Wikigame conducted on Wikipedia's link network – can contribute to computing semantic relatedness of concepts using the introduced concept co-occurrence (cf. Section 3.3) in Section 4.1. In Section 4.2 we compare the results to those obtained from several baseline corpora to show the additional benefit of human navigational paths.

4.1. Contribution of navigational paths to semantic relatedness

To show the usefulness of human navigational paths for calculating semantic relatedness we conduct our experimental steps as described in Section 3.3 where we not only use sliding windows of varying size *k* but also the principle that all concepts in a path co-occur with all other present concepts in the path on the corpus of all available paths taken from "TheWikiGame" – which we denote as a "none" window size. One can think of the "none" window size as a size that is always exactly as long as the path.

Table 3 presents the evaluation results for varying window sizes. We report the number of pairs (shown in column *#pairs*), for which we can calculate a semantic relatedness score (stated in column *ws353*). The reason why one can not

may get to judge the actual results.

always evaluate against each single pair of concepts is that there might not be co-occurrence information available for concepts of pairs using a specific window size – i.e., generally the larger the window size, the more pair scores we can calculate. A first observation is that the method of letting all concepts in a path co-occur with all other concepts in the path denoted as "none" performs worse than some specific sliding window sizes denoted in the table. This strengthens our assumption that the distance between two concepts in a path is crucial for calculating precise semantic relatedness scores as pointed out in Section 3.3. Furthermore, we can see that the best accuracy can be achieved using a window size of $k = 3$ or $k = 4$. Hence, letting the surrounding two or three concepts ($k-1$) given a concept in a path co-occur with the concept seems to be the most precise co-occurrence representation for determining the semantic relatedness between concepts in our corpus of human navigational paths. Interestingly, this observation correlates with the distance often applied in graph based methods for word sense disambiguation, as reported in Navigli & Lapata (2010).

To investigate the usefulness of our approach of reporting results obtained from evaluating the scores of all possible WordSimilarity-353 pairs for a specific window size or corpus, we also repeat the experiments by using all 353 word pairs and setting the relatedness scores to zero if we can not cover a pair as this is frequently done in related work (see the last column in Table 4). However, this method introduces high negative bias to the results as we observe that not surprisingly, those window sizes or corpora perform better that can simply cover

Table 3: Semantic relatedness calculated on human navigational paths. Our corpus consists of all available Wikigame paths where different window sizes ($2 \leqslant k \leqslant 5$) as well as the principle that all concepts in a path co-occur with all other concepts in the path denoted by "none" were evaluated against the WordSimilarity-353 golden standard by calculating the Spearman's rank correlation coefficient between the produced rankings of each method and the ones of the WordSimilarity-353 gold standard.

window size	#pairs	ws353
none	299	0.649
2	236	0.638
3	275	0.709
4	286	0.718
5	293	0.690

more WordSimilarity-353 pairs. We also calculate statistical significance tests between the dependent Spearman's rank correlation coefficients produced by different window sizes for this evaluation method using a one-tailed hypothesis test for assessing the difference between two paired correlations (Steiger, 1980). While the results indicate no statistic significant differences between window sizes 3 to 5 it is clearly visible that we would e.g., prefer an window size of 5 over "none" (the p-value $5.2 * 10^{-5}$ of the test is below the significance level of 0.05). Summarized, this evaluation represents a pessimistic evaluation compared to our optimistic one which only evaluates against possible word pairs, as it is hard to judge whether better accuracy is based on more precise calculations of semantic relatedness or simply more well defined term pairs. To further strengthen our evaluation approach we limit the evaluation in Table 3 to those pairs available throughout all window sizes (236 pairs) – see fifth column in Table 4 – and we can observe the exact same trend as our optimistic evaluation approach showed. Finally, we also sample 100 random pairs 100 times and average the results again showing in the fourth column of Table 4 that the best accuracy can be achieved using a window size of $k = 3$ or $k = 4$ and making a strong point for our evaluation approach. This agrees with similar observations by Ito et al. (2008) when evaluating against different subsets of WordSimilarity-353 pairs that the trend of accuracy always stays the same. Also, Milne & Witten (2008) pick up on this point as they directly

Table 4: Semantic relatedness accuracy calculated in similar fashion as for Table 3. This time, we report a variety of different evaluation approaches: (a) "possible pairs" reports the same results as in Table 3 and represent our optimistic evaluation, (b) "100 pairs" reports accuracy by sampling 100 word pairs 100 times and averaging the results, (c) corresponds to the accuracy by using only those word pairs that can successfully be determined for all windows sizes and (d) "all pairs" fills in zero semantic relatedness scores for word pairs for which no score can be calculated and represents the pessimistic evaluation. The observations illustrate the usefulness of our proposed "possible pairs" method.

window size	#pairs	possible pairs	100 pairs	236 pairs	all pairs
none	299	0.649	0.630	0.632	0.548
2	236	0.638	0.633	0.638	0.560
3	275	0.709	0.692	0.694	0.588
4	286	0.718	0.697	0.695	0.587
5	293	0.690	0.690	0.692	0.589

show that as they only include well-defined term pairs to their evaluation, they can achieve the appropriate results.

As the goal of this work is not to achieve the best possible semantic relatedness scores in comparison to related work techniques, but rather to identify whether and if so, human navigational paths can contribute to this task and to find the most appropriate window size and path corpus, we only report results obtained from applying our optimistic evaluation procedure which evaluates the scores of all possible WordSimilarity-353 pairs for a specific corpus. Note that we will also only cover a very small amount of pairs later on for our sampling strategies which makes the other evaluation methods not applicable – i.e., only using the same intersection of pairs for all methods would limit the gold standard tremendously (max. 30 pairs) and using all pairs by filling in zeros for missing word pairs would have high negative influences on methods that can only cover a small amount of pairs due to lack of data. This choice is based on abovementioned investigations and observations and gives us a logic way to evaluate our work. Due to tractability, we focus on window size $k = 3$ for the rest of this paper[15].

Table 3 demonstrates that human navigational paths contain information relevant for calculating semantic relatedness between concepts by exhibiting high quality relatedness evaluated against WordSimilarity-353. We investigate the additional benefit of the paths at hand to several baseline corpora next.

4.2. Additional benefit of navigational paths

As our human navigational paths of “TheWikiGame” are basically subsets of the underlying topological link network we need to investigate whether the observed effects are based on human intuitions and patterns while navigating or if automatic extractions of paths from the link network can produce similar or even better results. By doing so we can also investigate which role the rich topological link network plays for calculating semantic relatedness on paths.

To get first insights, we highlight basic properties of the Wikipedia link structure that we have studied, and the corresponding navigational paths that we have obtained in Figure 2. The figure contrasts the degrees of nodes in a subset of Wikipedia with the number of clicks on these nodes in a baseline random walk and in human navigational paths. As we see, the number of clicks on nodes from

[15] Note that a window size of $k = 4$ is just by a small margin more precise than a window size of $k = 3$ and the reason for only reporting results for $k = 3$ is based on faster runtime and better possibilities for interprating the results or looking into fingerprints. Nevertheless, we have also conducted further experiments by using a windows size of $k = 4$ which exhibit similar patterns.

human navigational paths differs significantly from (i) the network topology and (ii) the clicks generated by a random walk. On the one hand, we can see that human navigation tends to focus on a few nodes more heavily than a random walk on the network topology would lead us to expect, while on the other hand, they seem to place less focus on a wider range of nodes[16]. As both random walk and human navigational paths are basically subsets of weighted links, we can see that the weights emerging from user's choices during the game differ from the weights produced by a random walk. Hence, we want to explore whether these differences resulting from actual human navigation in information networks provide additional value for calculating semantic relatedness in comparison to navigation done by an automatic agent. To do so, we compare the corpus of paths from "TheWikiGame" with several baseline corpora which we introduce in the following sections. Finally, we present the results in Section 4.2.5.

4.2.1. Topological neighbor paths

A rather simple baseline for comparison consists of artificial sub-paths taken from Wikipedia's link network limited to concepts available in our Wikigame dataset (see Section 3.4.2). Given Wikipedia's topological (limited) link graph $\mathbb{W}_{wg} = (V_{wg}, E_{wg})$ with vertices V_{wg} and directed edges $E_{wg} = \{(v, w)|v, w \in V_{wg}\}$, we generate all possible paths of length three, where every node still lies in V_{wg}. This gives us the following set of paths $\mathbb{P}_{tb} = \{(u, v, w)|(u, v), (v, w) \in E_{wg} \cap V_{wg} \times V_{wg}\} \subset \mathbb{P}$.

The reason for choosing paths with the length three for this topological baseline corpus is that we focus on a window size of $k = 3$ – i.e., a concept co-occurs with the neighboring $k - 1 = 2$ concepts in a path – throughout this work (see Section 4). Hence, with this corpus of artificial paths we can calculate all possible co-occurrences between concepts in a window of size $k = 3$. For this baseline, we will not only report results based on co-occurrence vectors with their respective co-occurrence counts, but also based on *binary vectors* – i.e., two concepts get a co-occurrence count of one if they appear in at least one single path of length three together – ignoring the number of co-occurrences and thus controlling the vast amount of artificial paths. This enables us to investigate the influence of the degree of concepts on the results – note that again the extracted corpus of paths is a weighted subset of the plain Wikipedia link structure where the weight is in-

[16] The *Kullback-Leibler divergence* (0.738) and the *Spearman rank correlation* (0.130) between the click distributions of a random walk vs. human navigation indicate that there is indeed a significant difference between the distributions.

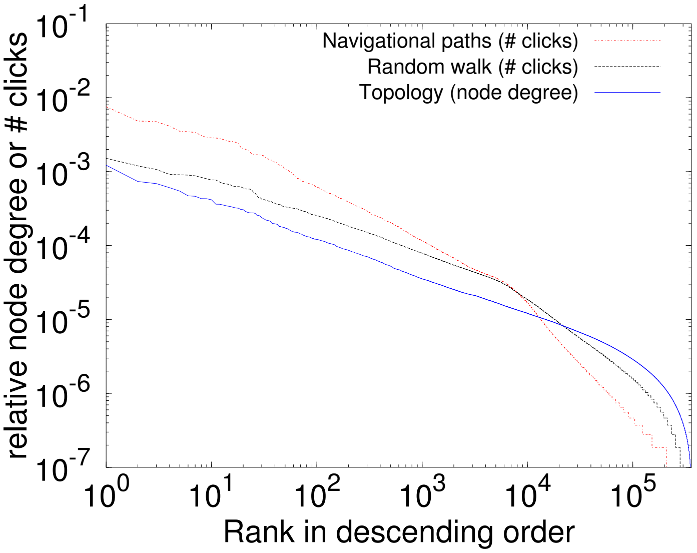


Figure 2: Properties of the Wikipedia link structure that we have studied and corresponding navigational paths that we have obtained. The figure compares the distribution of node degrees of the underlying Wikipedia topology (blue solid line) with the relative click frequency on the same set of nodes obtained from a random walk (black dashed line) and from navigational paths obtained from the Wikigame (red dotted line). The ranks on the x-axis are based on the corresponding node degree or #clicks for the corresponding node in descending order – e.g., the node with rank 1 has the highest degree.

fluenced by the degree of each node (e.g., a node with an out-degree of 4 is more likely to get higher co-occurrence counts than a node with an out-degree of 1).

4.2.2. Biased random walk paths

We aim to compare the usefulness of human navigational paths to artificial paths (e.g., produced by an algorithm) as another kind of baseline. Therefore, we perform a biased random walk through Wikipedia's underlying plain topological link structure preserving some of the structural information taken from our Wikigame paths. For each path, we select the start node and initialize a random walk on Wikipedia's link network limited to concepts available in the Wikigame. The random walk then walks freely through this network by choosing a random outlink available for a concept. The walk stops when the similar path length as the corresponding Wikigame path is reached. By doing so we end up with a corpus of paths that approximately has the same number of visited pages as “TheWikiGame” corpus, but exhibits dissimilar link weights. If the walker reaches a concept with no out-link, it goes back one position and tries another path. The relative concept click frequency of the resulting paths can be seen in Figure 2. We call the resulting set of random paths $\mathbb{P}_{random}$.

4.2.3. Permuted Wikigame paths

To understand how important the underlying link structure is for the task of calculating semantic relatedness on navigational paths and also to explore how much impact the sequence of concepts in a human navigational path has, we create so-called *permuted paths*. In these paths, we are still leaving the position of a concept in a path intact, but swap it with a node on the same position of another path and by doing so we detach the node with preceding and succeeding nodes of the path. For a given path $p = (v_1, \ldots, v_n) \in \mathbb{P}$, we randomly choose another path $q = (w_1, \ldots, w_m)$ and randomly swap a node in p with the corresponding node at the same position in q. We receive two new paths $p' = (v_1, \ldots, w_i, \ldots, v_n)$ and $q' = (w_1, \ldots, v_i, \ldots, w_m)$ where we lose the semantic information around the newly inserted node. Again, we preserve as much structural information as possible of our game paths while randomizing the semantic related information. We call the resulting path set $\mathbb{P}_{permuted}$. It is important to note in this scenario, nodes might not be linked from their predecessor or to their successor on the underlying Wikipedia topology. These newly created paths are called $\mathbb{P}_{permuted}$ and contain exactly as many paths as $\mathbb{P}$.

4.2.4. Swapped Wikigame paths

The purpose behind this method is to keep the link structure of Wikipedia intact but to swap out parts of a supposedly meaningful path with parts of another path. Our method works as described in the following: For a given path $p = (v_1, \ldots, v_{i-1}, v_{mid}, v_{i+1}, \ldots, v_n)$, we select another path $q = (w_1, \ldots, w_{j-1}, v_{mid}, w_{j+1} \ldots, w_m)$ with maybe a different length, but with the property that the node v_{mid} is in the middle of both paths. We cut both paths in half and exchange the back part of p with the one of q in such a way that we receive the new paths $p\prime = (v_1, \ldots, v_{i-1}, v_{mid}, w_{j+1}, \ldots, w_m)$ and $q\prime = (w_1, \ldots, w_{i-1}, v_{mid}, v_{j+1}, \ldots, v_m)$. The newly generated paths are called $\mathbb{P}_{swap}$ and contain exactly as many paths as $\mathbb{P}$.

4.2.5. Results

Table 5 presents the results using a window size of $k = 3$ with all available Wikigame paths and our baseline corpora as described above. In column *#paths* one can see the number of paths available for each corpus and in column *length* the total accumulated length of all paths in the corpus. Finally, in column *#pairs* we can see the number of pairs of the WordSimilarity-353 dataset where we can successfully calculate the semantic relatedness measures and the final *Spearman rank correlation* to WordSimilarity-353 is shown in column *ws353*. Further insights from these investigations are discussed next.

Wikipedia topology alone is useful: We know from other semantic analysis methods, that the Wikipedia topology alone provides useful information. For confirmation, we evaluated the scores obtained from our *Permuted Wikigame paths corpus*. The corresponding results confirm that we lose semantic preciseness when ignoring the original link and navigation structure. Keeping the original structure intact, but swapping parts of the paths – see *Swapped Wikigame paths* and the corresponding description above – we can see that the original navigation by a user has a high impact on the achieved accuracy, but that we can still achieve reasonable results by leaving the underlying link structure and partly navigational patterns intact. We can also see that *Biased random walk paths* perform similar to our *Topological neighbor paths corpus*. This is not surprising, as the random walks freely navigate the topological link network, even though they are biased towards a specific path length and are initialized by a given start node.

Human navigation paths improve results: A first observation is that the Wikigame path results outperform the baselines by a relevant margin – for example, it outperforms the best baseline method *Swapped Wikigame paths* by 0.041

Table 5: Comparison of semantic relatedness calculations using a window size of $k = 3$ evaluated against WordSimilarity-353 on all Wikigame paths with several baseline corpora.

Corpus	#paths	#pairs	ws353
All Wikigame paths $\mathbb{P}$	1,799,015	275	0.709
Topological neighbor paths $\mathbb{P}_{tb}$	6,042,578,644	308	0.659
Topological neighbor paths $\mathbb{P}_{tb}$ binary	6,042,578,644	308	0.485
Permuted Wikigame paths $\mathbb{P}_{permuted}$	1,799,015	292	0.381
Swapped Wikigame paths $\mathbb{P}_{swap}$	1,799,015	273	0.668
Biased random walk paths $\mathbb{P}_{random}$	1,797,326	274	0.660

(0.709 vs. 0.668). When looking at the *Topological neighbor paths corpus*, we can also see that Wikipedia's inherent link structure already can be used as a powerful resource for calculating semantic relatedness using our methodology. In order to see how the number of co-occurrences between concepts influences the semantic relatedness, we have also performed an analysis on the same corpus of topological neighbor paths, but this time we do not count how often two concepts co-occur, but only represent the co-occurrence state with a binary value – we can refer to this as the *plain structure*. We can now see that the accuracy evaluated against WordSimilarity-353 drops by a significant amount (from 0.659 to 0.485) indicating that the number of co-occurrences between concepts effects our method. We can observe from this that the weighting of links in a path corpus has high impact on the accuracy we can achieve.

With this initial exploration, we can conclude that human dynamic navigational paths on Wikipedia can contribute to computing semantic relatedness, but they are based on an already powerful network topology. The weighting provided by users' choice during navigation exhibits the most precise information for determing semantic relatedness between concepts. Next, we want to identify what kind of navigational paths are most useful for that task.

5. Path selection experiments

Human navigational paths can be characterized along many dimensions. For example, there exist *successful paths* where users were able to successfully reach the specified target nodes, while on *unsuccessful paths* users could not reach their goal. Other path characteristics may mostly move along high degree (vs. low degree) nodes. Figure 3 shows the distribution of path lengths in all paths (black line), only in successful paths (red line) and only in unsuccessful paths (blue line). Only looking at such path length distributions, we can already see that such distinct path types exhibit different features. We want to explore these differences and investigate their usefulness for the task of calculating semantic relatedness, e.g., investigate whether a subset of only successful paths is more useful than a subset of only unsuccessful paths.

This gives rise to a number of interesting questions related to different navigational paths, such as (a) *Are all navigational paths equally useful for computing semantic relatedness?* and (b) *If some navigational paths are more useful, what are the characteristics of these paths and how can they be exploited?* To analyze these and other questions, we begin our investigations by taking the corpus of all *successful paths* (which is the smaller set) and extract a random subset of *unsuccessful paths* of equal size, containing the same number of visited pages.

Similar to Section 4, we use a window size of $k = 3$ for our co-occurrence calculation and evaluate the relatedness scores against WordSimilarity-353; the results can be seen in Table 6. From that table, we see that a smaller subset of our corpus of all Wikigame paths $\mathbb{P}$ can perform remarkably well – compare with Table 5. Somewhat surprisingly, we see that a corpus of *unsuccessful paths* performs better than a corpus of *successful paths* with the same total number of visited concepts. A possible explanation for this behavior is that unsuccessful paths contain the behavior of mostly inexperienced users who try to follow nodes whose meanings are very close and hence, remain on a narrow semantic field which may also lose them the game. On the other hand, successful players might navigate through more distant concepts or very central concepts like "United States" which are common strategies for winning a game. Further investigations are necessary in order to explain this behavior in detail, which is not in the scope of this work.

Regardless the exact explanation of this behavior, the results suggest that subsets of paths with specific characteristics yield different results. This leads to the idea of investigating whether smaller sets of paths according to specific path characteristics can perform similarly or even more precise in regard to our relatedness calculations on the whole set of paths. In the following section, we will explore this by conducting different path selection experiments.

5.1. Characteristics of Paths

We introduce several measures $m : \mathbb{P} \rightarrow \mathbb{R}_0^+$ to characterize any path p in our corpus of paths $\mathbb{P}$. Each distinct measure makes use of a path characteristic, depending on the visited nodes, which actually characterize the path. The resulting measures will be subsequently used in section 5.2 to create path selections.

In the following, we will elaborate each of the different measures in greater detail. Let $p \in \mathbb{P}$ be an arbitrary path represented by the sequence of nodes

Table 6: Comparison of semantic relatedness calculations using a window size of $k = 3$ evaluated against WordSimilarity-353 on all Wikigame paths with several baseline corpora.

Corpus	#paths	length	#pairs	ws353
Successful Wikigame paths	653081	4116879	230	0.636
Unsuccessful Wikigame paths	710374	4116879	257	0.683

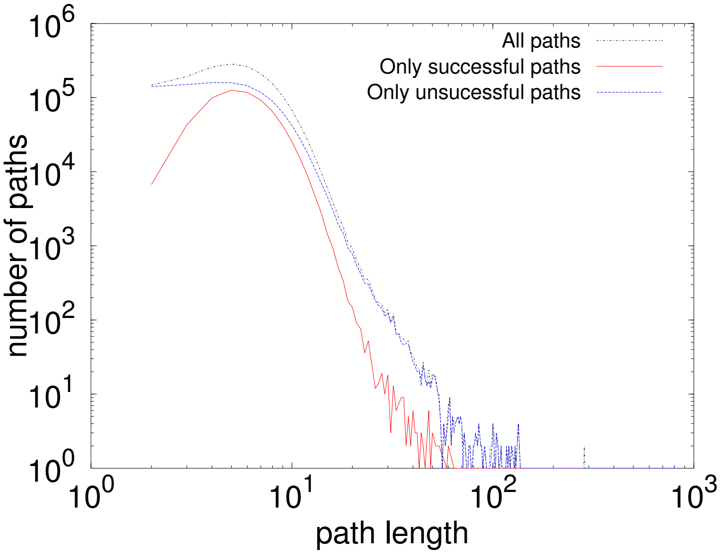


Figure 3: Illustration of the distribution of path lengths in all human navigation paths (black dotted line), only in successful paths (red solid line) and in unsuccessful paths (blue dashed line).

$(v_1, \ldots, v_n)$.

In- and outdegree. For a path p, we determine the in- and outdegree for each concept v_i in p derived from Wikipedia's complete topological link network. The idea behind this characteristic is to differentiate hubs and strongly connected concepts from dead ends and rather weakly connected concepts. The measure is calculated as ($m_{outdegree}(p)$ is defined analogously):

$$m_{indegree}(p) = \frac{1}{len(p)} \sum_{k=1}^{n} indegree(v_k).$$

Ratio. This measure represents a ratio of in- and outdegree for each node in a corpus of paths smoothed by the square root of the indegree (see (Trattner et al., 2012)). This characteristic is motivated by the notion that a page with e.g., 200 inlinks and 100 outlinks should be more important than a page with two inlinks and one outlink. If the outdegree for a node is zero, we set the ratio to zero as well. $ratio(v)$ is calculated in the following way for a node v:

$$ratio(v) = \frac{indegree(v)}{outdegree(v)} \cdot \sqrt{indegree(v)}.$$

Thus, the value of a path p is determined by

$$m_{ratio}(p) = \frac{1}{len(p)} \sum_{k=1}^{n} ratio(v_k).$$

TF-IDF. Interpreting a path as a document and the concepts present in a path as terms, we use the well known *tf-idf* scores (cf. (Salton & Buckley, 1988)) of each node in a path as a further characteristic. The idea behind this characteristic is that we can identify paths that include many concepts that are very important for the individual path compared to all other paths in the corresponding corpus. Hence, for each path p, we again take the mean of all tf-idf values in the path:

$$m_{tfidf}(p) = \frac{1}{len(p)} \sum_{k=1}^{n} tfidf(v_k).$$

Length. Finally, we use the length of a path p – i.e., the number of concepts visited in a path – as a last characteristic:

$$m_{length}(p) = len(p).$$

Our motivation for taking the length of a path as a characteristic is the notion that longer paths potentially contain more information because of more co-occurrences between concepts of the paths. Furthermore, we could observe in Figure 3 different path length characteristics for different types of Wikigame paths, which is interesting to investigate in greater detail.

5.2. Path selection strategies

Based on the characteristics described in Section 5.1 we now select smaller sets of paths according to abovementioned path characteristics. We investigate whether the relative performance of reduced corpora of paths $\mathbb{P}_m$, based on the accuracy of our relatedness scores, increases or decreases, compared to the performance of our complete set of paths $\mathbb{P}$, in analogy to Koerner et al. (2010).

For each characteristic, we calculate ten subsets of increasing size where the tenth subset corresponds to the set of all available Wikigame paths. The sizes of our subsets are calculated by the number of visited nodes inside the subset. If we consider the sum of all nodes $node\_sum = \sum_{p \in \mathbb{P}} len(p)$, a path selection of e.g., 10% does not necessarily contain $0.1 \cdot |\mathbb{P}|$, but rather $0.1 \cdot node\_sum$. More formally, we can express it as follows: Consider an ordered list $l_m = (p_1, \ldots, p_n)$ of paths, generated by a measure m. A selected subset $\mathbb{P}_x^m$ of size x for measure m can be expressed as:

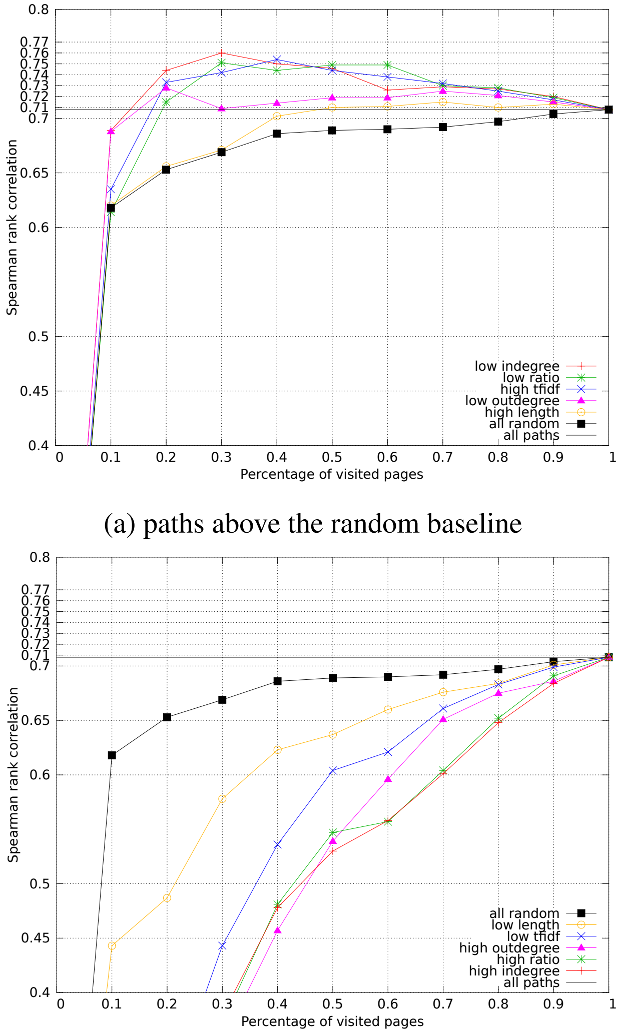
$$\mathbb{P}_x^m = \left\{ p_k | k = \max \left\{ s | \sum_{j=1}^{s} len(p_j) \leqslant \frac{x}{100} \cdot node\_sum \right\} \right\}. \tag{2}$$

Thus, a potential path selection with very long paths consists of fewer actual paths than a selection with mostly short paths, but both sets contain roughly the *same amount of visited nodes*. This renders the selection process more fair than just pure path counting as it enables us to fairly compare two corpora of the same size selected based on different path characteristics. Each selection process generated subsets consist of $x = \{10, 20, \ldots, 90\}\%$ of all visited pages. By proceeding with this selection process, the first subset – i.e., the 10% subset – consists of paths with the lowest measures for a corresponding characteristic – e.g., paths with the lowest mean indegree. Furthermore, we also revert the ordered list l_m in order to get a ranking $l_m^{rev} = (v_n, \ldots, v_1)$ where the small subsets contain paths with higher measures for a specific characteristic – e.g., paths with the highest mean indegree. After the generation of the path ordering lists and the path selection process described above, we run our semantic evaluation for each of these subsets.

Furthermore, we create a baseline for each individual split to learn whether the distinct accuracy results are genuinely dependent on the corresponding path selection process based on several characteristics. We shuffle the corpus of paths independently and randomly ten times in order to remove the original ordering in the complete set of paths. For each of these ten independent shuffles, we extract subsets according to the selection process described above. We end up with ten selections for each subset containing $x = \{10, 20, \ldots, 90\}\%$ of the visited pages. Finally, we perform our semantic analysis and evaluate the results accordingly for each selection and subset. We average the results for each subset based on the sum of selections for the corresponding subset and report the results in the following section; we will refer to this baseline as *random baseline*.

5.3. Results

In Figure 4 we present the results obtained from our individual sub-corpora of navigational Wikigame paths using our selection strategies pointed out in Section 5.2 based on characteristics of paths – or to be precise: characteristics of concepts inside paths averaged for each path – described in Section 5.1. Figure 4a illustrates selections where we can achieve better accuracy – i.e., Spearman rank correlation evaluated against WordSimilarity-353 – than using random selections



(a) paths above the random baseline

(b) paths below the random baseline

Figure 4: Semantic relatedness calculated on different path selections. Larger values on the y-axis correspond to higher Spearman rank correlation with the WordSimilarity-353 dataset. The black horizontal line depicts the result for the entire set of paths. Figures 4a and 4b show the results of different selection strategies. Figure 4a shows selection results with better-than-random performance while Figure 4b shows results with worse-than-random performance. In Figure 4a, we can see that only a small subset of 30% low indegree paths produces more precise semantics than the whole path corpus $\mathbb{P}$ would (scoring a rank correlation of 0.760 to the WordSimilarity-353 dataset). Paths characterized by low in- and outdegree always perform better than a random baseline, while their counterparts, starting from high degrees, perform significantly worse. Similar patterns can be observed when selecting paths according to their tf-idf values.

of all Wikigame paths – i.e., *random baseline* (black line, ■) – while Figure 4b shows selections performing worse. The horizontal black line with a Spearman rank correlation of 0.709 shows the results achieved when taking a corpus of all Wikigame paths (see also Table 3). For all selections we use a window size of $k = 3$ for our co-occurrence and subsequent semantic relatedness calculations. Our key findings are discussed next.

Intelligent path selection improves semantic relatedness. A first observation when looking at Figure 4a is that smaller random path selections do not lead to a similar or better accuracy (black line, ■), but that we indeed can find smaller corpora of navigational paths – selected on several characteristics – that perform equally or better than the complete corpus of Wikigame paths (that reaches an accuracy of 0.709). By incrementally adding paths with the lowest average indegree of their concepts, we can achieve the highest Spearman rank correlation with a sub-corpus of only 30% of all Wikigame paths (red line, $+$). The respective accuracy of 0.760 outperforms the accuracy of the whole Wikigame corpus by about 5% while covering less than a third of all visited pages in the complete corpus. Contrary, we can see in Figure 4b that a reverse accumulation of paths, beginning with those having a high average indegree (red line, $+$), leads to much worse accuracy compared with the random baseline and as well as with the accuracy of the complete corpus. A possible explanation for this is that low indegree nodes represent concepts that do not seem to be hubs nor exceptionally abstract concepts in comparison to high indegree nodes. Also, high indegree concepts may have much more co-occurrence counts with several other concepts while low indegree concepts may only have co-occurrence connections to a few very specific concepts (even when looking at a window size of $k = 3$). Hence, the co-occurrence vectors may be sparser, but more precise and this may enable us to calculate more accurate semantic relatedness scores. If we look deeper into the paths included in our selection corpora we can see that paths with the highest average indegree all include the concept *United_States* which is on the one hand, the most central concept in Wikipedia's topological link network, and on the other hand, also by far the most often navigated concept in our Wikigame paths. Hence, this concept co-occurs with many others and is no suitable descriptor for determining the semantic relatedness between concepts while paths with the lowest average indegree contain more variety and also more descriptive co-occurrences. To summarize: **Small selections of low indegree paths exhibit more fine-grained and precise semantics than the set of all paths.**

To give an example we illustrate in Figure 5 the concept co-occurrence vectors for the concepts *Vodka*, *Brandy* and *Bread* on the one hand, using our best

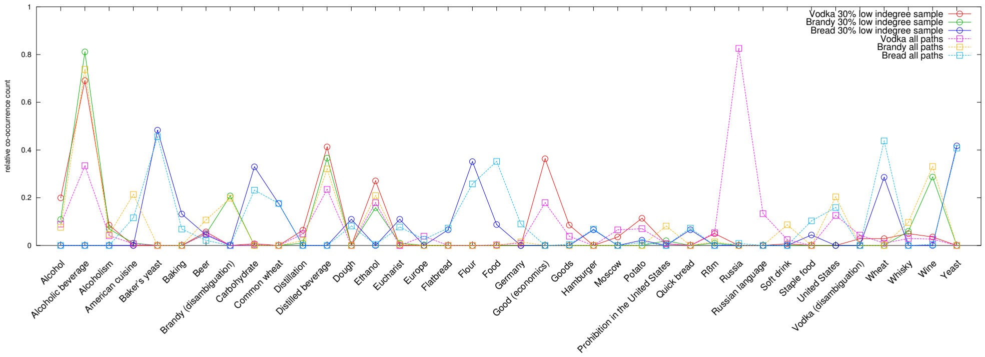


Figure 5: Semantic “fingerprints” for the concepts *Vodka*, co-occurrence count of fifteen to the corresponding concept. All counts are normalized by the L2 norm of the vector and fingerprints for a 30% low indegree selection (solid lines) and the full set of paths (dashed lines) are shown. The 30 % low indegree selection exhibits more fine-grained and precise semantics than the set of all paths.

overall performing corpus of 30% low indegree paths (solid lines, ◯) and on the other hand, deriving the information from the all path corpus (dashed lines, ■). For visualization purposes the vectors are reduced in dimensionality by only representing co-occurrences to concepts where at least one vector exhibits a count of larger than 15. Furthermore, all counts are normalized by the L2 norm of the complete vector. In Figure 5 we can see that the concepts *Alcoholic beverage*, *Distilled beverage* and *Ethanol* exhibit similar peaks for the concepts *Vodka* (red solid line, ◯) and *Brandy* (green solid line, ◯) for the corpus of 30% low indegree paths, while having only few diverse peaks. We can observe that these common peaks contribute a lot to the high cosine similarity of 0.8043 that we can compute with this subset for the corresponding concept pair. This score agrees extremely well with the human score of 8.13 present in the WordSimilarity-353 dataset. In contrast, we can see that there are only a few similar normalized co-occurrences for the concepts *Vodka* (pink dashed line, ■) and *Brandy* (orange dashed line, ■) using the corpus of all paths and that the concept *Russia* exhibits a large diversity regarding the co-occurrence patterns for both concepts negatively influencing the relatedness score resulting in only 0.4205. The co-occurrence vectors for the concept *Bread* show for both corpora – i.e., 30% low indegree paths (blue solid line, ◯) and all paths (turquoise dashed line, ■) – no common peaks to both other concepts resulting in extremely low relatedness scores. We can see from this, that our selection of low indegree paths exhibits much more fine-grained patterns for the concept pair *Vodka* and *Brandy* reaching also a higher relatedness score than our corpus of all paths by still keeping low scores for concept pairs, that are not semantically related.

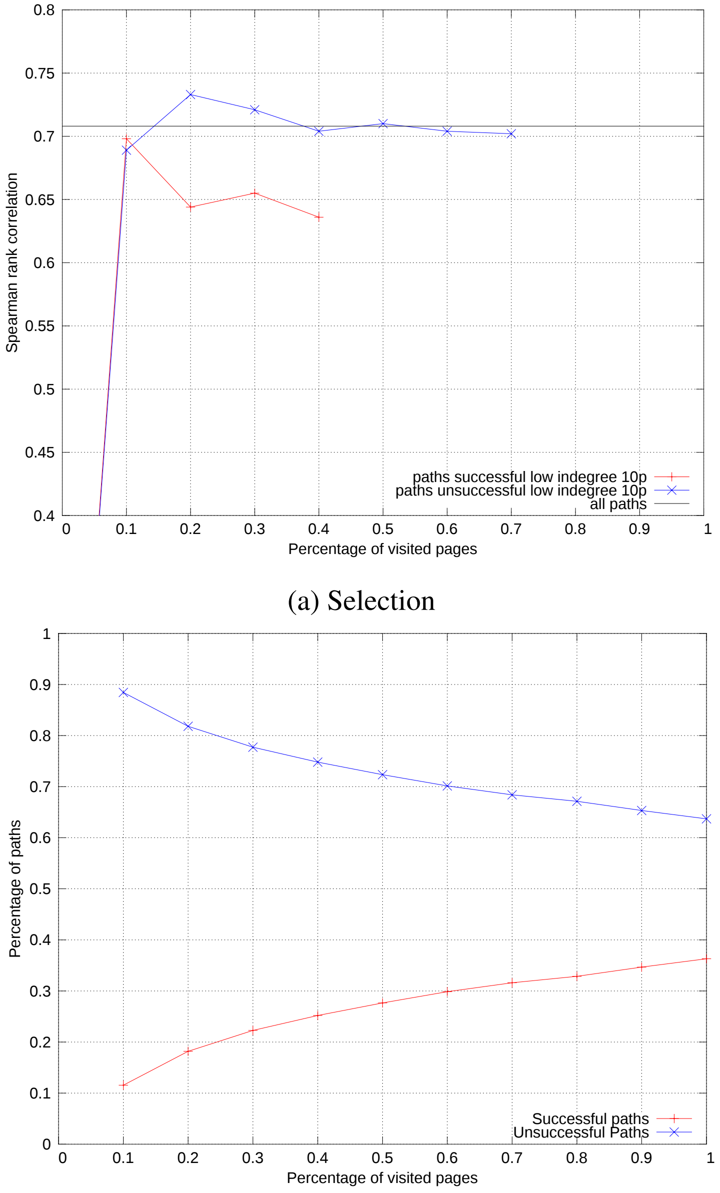
Other degree based selection strategies and corpus based characteristics (e.g., tf-idf) can also improve accuracy. Similar observations as above can be seen by selecting according to the average outdegree of paths starting with the lowest value depicted in Figure 4a (pink line, ▲). Smaller selections can outperform the corpus of all paths, but we can not achieve as good results as with our 30% selection of low indegree paths. Again, the opposite occurs for the reverse selection of paths starting with those having a high outdegree shown in Figure 4b (pink line, ▲) – i.e., all selections perform worse than the baseline and the complete corpus. Selections based on the average *ratio* of paths (green line, ✳) not surprisingly show similar patterns as the selection according to in- and out-degree, but indicate that a selection according to the average indegree of paths can achieve higher accuracy than using a combined measure. Selection strategies based on the *tf-idf* values of nodes inside paths indicate that we can strongly outperform the baseline and the target accuracy of a corpus of all paths for several sub-corpora incrementally adding paths with a high average tf-idf value shown in Figure 4a (blue line, ×). Contrary, selecting paths with low tf-idf scores never reaches the accuracy of the random baseline as we can see in Figure 4b (blue line, ×). Low average tf-idf valued paths exhibit similar patterns than those with a low average indegree. The difference though is that this measure is only corpus dependent and ignores characteristics of the underlying topological link network and this may exhibit advantages for specific scenarios. Finally, we can see from both illustrations in Figure 4 that a selection according to the length of paths (orange line, ○) produces just three sub-corpora of paths – i.e., 70% to 90% selections of longest paths – that can slightly outperform the corpus of all paths.

A combination of successful and unsuccessful paths produces more precise semantics than using unsuccessful paths only. Our initial experiments showed that a corpus of unsuccessful paths outperforms a corpus of successful paths in regard to the accuracy of our semantic relatedness scores (see Table 6). Now that we know that a path corpus with lower indegree paths works better one possible reason for the better performance of unsuccessful paths might be that the average indegree of unsuccessful paths is lower as the average indegree of successul paths as we have investigated. However, with the observation that there are more intelligent ways of selecting a corpus of paths accordingly (e.g., by selecting low indegree paths), the question arises if we can furthermore improve the preciseness of semantic relatedness calculation by performing a similar selection just on the corpus of unsuccessful paths. To this end, we use our best performing characteristic measure – namely the *indegree* – and select in the typical way sub-corpora of unsuccessful paths starting with those having the lowest mean indegree. We do the same selection for successful paths to be able to compare both subsets. Again, we accumulate the number of paths in a selection towards the total number of visited nodes of the corpus of all paths; we end up with more selections for unsuccessful paths than for successful paths as we have a larger fraction of unsuccessful paths.

In Figure 6a we identify that we can outperform the horizontal black solid line indicating the accuracy obtained from a corpus of all Wikigame paths. The best results can be achieved by using a 20% split of only unsuccessful paths (blue solid line). While this accuracy of 0.733 outperforms the whole set of all paths, we still get a better result by selecting the whole corpus in a similar fashion as depicted in Figure 4a, where we could reach an accuracy of 0.760. When we now look deeper into the subsets of low indegree based selections calculated for the complete dataset, we see that around 25% of the paths inside the best performing 30% low indegree sub-corpus (selected on all paths) are successful paths (see Figure 6b). While unsuccessful paths tend to exhibit characteristics that make them more useful for computing semantic relatedness, we find that overall a combination of successful and unsuccessful paths produces the best results. The results also suggest that other characteristics such as the indegree and not success are better suited for selecting good subsets when performed on the whole set of paths.

Evaluating against other gold standard datasets confirms our observations. Throughout this section we have only used the *WordSimilarity-353* dataset as a gold standard for our evaluations. The reason for this choice was that it is a widely used gold standard for evaluating semantic relatedness scores against human judgements. Nevertheless, there also exist other prominent datasets similar to WordSimilarity-353: (a) the *Miller Charles gold standard* (Miller & Charles, 1991) (30 overall word pairs) and (b) the *Rubenstein Goodenough gold standard* (Rubenstein & Goodenough, 1965) (65 overall pairs). In order to triangulate our observations, we conducted the same experiments on both datasets by mapping words to concepts manually and calculating Spearman rank correlation. Again, we make our mappings available online[17]. The results for both gold standards are illustrated in Figure 7. Again, we can clearly see that we can outperform the accuracy of the complete set of paths by sampling smaller sets affirming the patterns

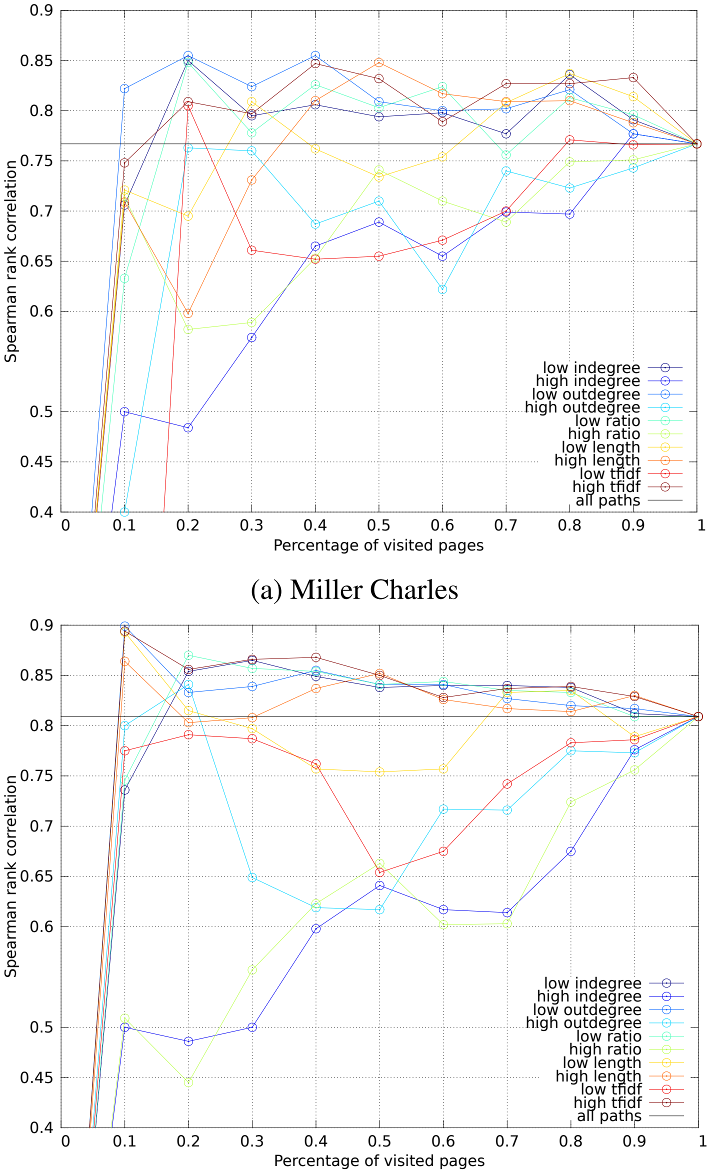
[17] http://www.philippsinger.info/wikisempaths.html



(a) Selection

(b) Proportion

Figure 6: Effect of successful / unsuccessful paths: (a) shows successful (red solid line) and unsuccessful paths (blue solid line) selected according to their average indegree starting with low indegree paths and their respective Spearman rank correlation evaluated against WordSimilarity-353. (b) shows the percentage of successful (red solid line) and unsuccessful paths (blue solid line) for our best performing selection of 30% low indegree paths (see Figure 4a). While path selection based on unsuccessful paths performs better than a selection of successful paths, we can see that overall a combination of successful and unsuccesful paths produces the most accurate results.



(a) Miller Charles

(b) Rubenstein Goodenough

Figure 7: Semantic relatedness evaluation for our sampling strategies evaluated against both the Miller Charles and Rubenstein Goodenough gold standards calculated and illustrated in similar fashion as in Figure 4. Again, we can see that specific samples can outperform the corpus of all Wikigame paths by excelling their corresponding Spearman rank correlations of 0.767 for the Miller Charles dataset and 0.809 for the Finkelstein Goodenough dataset.

observed in these experiments. This indicates that such sampling strategies can help us to remove paths with some kind of *semantic noise* by e.g., ignoring paths with a high average indegree of their visited concepts. By doing so, we can produce more precise semantics out of navigational path data. However, we need to take the results for these two additional evaluations with caution, as both gold standards are very limited in their number of word pairs they cover. We also can only capture at maximum 21 pairs for the Miller Charles and 40 pairs for the Rubenstein Goodenough dataset, while some samples can only cover a very low amount of pairs. Hence, this may also give rise to the slight unstable results in Figure 7 as sometimes the samples might simply capture very well-defined concept pairs while leaving others out. Contrary, this is not the case for the WordSimilarity-353 dataset where much more word pairs are available and where we can also cover much more pairs for all sub-samples.

6. Discussion and conclusions

To the best of our knowledge, this work represents the largest and most comprehensive effort to study semantics in human navigational paths to date. We (i) systematically evaluated information on ∼*1.8 million human navigation paths* captured via a semi-controlled navigation experiment against baselines that use the Wikipedia topology only or alternate the human navigational paths at hand and we (ii) evaluated the results against common reference datasets of relatedness. The main contributions of our work are the following. (1) Our experiments further indicate that such human navigational paths can represent a viable source for calculating semantic relatedness between concepts in information networks. (2) We show that semantic relatedness calculated based on human navigational data may be more precise than semantic relatedness computed from Wikipedia's link structure alone and (3) we find that not all navigational paths are equally useful. Intelligent selection of navigational paths based on path characteristics can improve accuracy.

If we compare our results to those obtained by previous works evaluated on the same full gold standard (see results from some well-known methods in Table 1) we can observe that we can match the accuracy of existing methods (our best score ends up at 0.76). Yet, there are obstacles in comparing the results to other methods directly. The main evaluation process of most of the related work remains a black box. Only slight adoptions to the Wikipedia dump used – e.g., by removing low degree concepts as ESA does – can already change the outcome tremendously. As the goal of this work is not to achieve the best performing method but rather detect signals in the data and show the usefulness of our approach we will not directly try to compare us with other works due to abovementioned reasons.

The method of leveraging human navigational paths using co-occurrence information presented throughout this work could also provide opportunities for improving existing content based methods in the sense of complementary information. For example, we could easily enrich existing co-occurrence based methods by interpolating the information extracted from human navigational paths. This would be a great way to incorporate pragmatic patterns to the content itself. In future, we want to concretely investigate the usefulness of such an approach by using navigational information by humans as additional signals for semantic relatedness for existing approaches.

A main limitation of this work is that we focus on human navigational paths derived from a game – namely "TheWikiGame". The game design itself may affect the structure of the paths and the resulting semantic relatedness scores. Some possible constraints of the game are: (a) a random choice of start and target nodes – hence, users also do target based navigation instead of pure exploration navigation, (b) users have a time constraint while navigating or (c) users tend to evolve strategies in order to win a game that may be counterproductive in terms of specifying semantic relatedness. Contrary, one could argue that real navigation more focuses on the goal of getting as much information as possible. One could also argue that such real human navigational data can even be more useful as humans may take more time for checking the current page and the next link would be chosen more accurately. They may also navigate on a more semantically narrow path. Nevertheless, the human navigational Wikigame paths present an abstraction of real user navigation in information networks and provide a further signal that such data can indeed be very useful for calculating semantic relatedness. In future we want to investigate human navigational paths in a less controlled navigational setting and investigate whether such paths can also contribute as much – or as hypothesized even better – as the data at hand indicates.

As mentioned throughout the work, our Wikigame paths are basically a subset of weighted links. Even though our results suggest that these paths can be more precise than artificial paths derived from Wikipedia's topological link network – note that these paths are again path sub-corpora of weighted links, where the weight is determined by an algorithm – we do not know if there might be a configuration of weights that leads to better results. Nevertheless, it is a complicated and not trivial task to automatically determine such a configuration of weights. As we can see from our experiments, human navigational paths seem to produce weighted link paths that can be very precise when calculating semantic relatedness. So, we may be able learn weighting configurations with the help of human navigational paths in order to automatically derive paths based on such weightings that may be even better than the human navigational paths themselves.

Our results are not limited by our evaluation approach as a) WordSimilarity-353 is an established gold standard that is frequently used to evaluate methods for computing semantic relatedness and b) our experiments with alternative gold standards for semantic relatedness have produced results exhibiting similar trends (cf. Section 5.3). However, we want to extend our evaluation approach in future by showing the usefulness of our method of computing semantic relatedness by using the output for several NLP tasks like word sense disambiguation, recommendation or text segmentation. Furthermore, we want to establish automatic disambiguation processes for our pipeline.

The findings of this work have interesting implications for future research: i) While our results focus on semantic relatedness, it appears plausible that other semantic tasks, such as hypo/hypernym detection can benefit from data about human navigational paths as well. For example, West & Leskovec (2012) have found that navigation in semi-controlled settings tends to consist of two phases where in an initial exploration phase more abstract concepts are sought out, while in a subsequent exploitation phase more specific semantic concepts are selected. This could be used in future methods to compute different levels of abstractedness for concepts based on their position in navigational paths. ii) While we have studied the usefulness of human paths in a semi-controlled navigation scenario, a natural next step would be to study less controlled navigational scenarios - such as actual human navigation paths - and their usefulness for computing semantic relatedness. None of our measures for modeling navigational paths is constrained to semi-controlled navigation scenarios, and they can all be applied to less controlled scenarios as well. iii) Our work makes a compelling argument for expanding the existing arsenal of data sources for calculating semantic relatedness. It suggests that in addition to data from textual or structural (link) sources, *usage* data - such as human navigational paths - could play a pivotal role in the future. Hence, we can envision that future methods for computing semantic relatedness might not produce objective scores for semantic relatedness, but *subjective* scores that take into account how concepts are used and perceived by large user populations via analyzing their aggregate navigation behavior.

6.1. Acknowledgements

We want to thank Alex Clemesha (TheWikiGame) for access to the TheWikiGame dataset and Denis Helic (TU Graz) for his cooperation. This work is in part funded by the FWF Austrian Science Fund Grant I677 as well as by the DFG through the PoSTS project.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* NAACL '09 (pp. 19–27). Stroudsburg, PA, USA: Association for Computational Linguistics.

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychon. Bull. Rev.*, *6*, 328–337.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th international joint conference on Artificial intelligence* IJCAI'03 (pp. 805–810). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA*, .

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), *Proceedings of the 7th International Conference on The Semantic Web* (pp. 615–631). Berlin, Heidelberg: Springer-Verlag volume 5318 of *ISWC '08*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, *20*, 116–131.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence* IJCAI '07 (pp. 1606–1611). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hassan, S., & Mihalcea, R. (2011). Semantic Relatedness Using Salient Semantic Analysis. In W. Burgard, & D. Roth (Eds.), *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* AAAI '11. AAAI Press.

Helic, D., Strohmaier, M., Trattner, C., Muhr, M., & Lerman, K. (2011). Pragmatic Evaluation of Folksonomies. In *Proceedings of the 20th International Conference on World wide web* WWW '11 (pp. 417–426). New York, NY, USA: ACM.

Hovy, E., Navigli, R., & Ponzetto, S. P. (2012). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, .

Ito, M., Nakayama, K., Hara, T., & Nishio, S. (2008). Association Thesaurus Construction Methods based on Link Co-occurrence Analysis for Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* CIKM '08 (pp. 817–826). New York, NY, USA: ACM.

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics* (pp. 19–33).

Koerner, C., Benz, D., Strohmaier, M., Hotho, A., & Stumme, G. (2010). Stop Thinking, Start Tagging - Tag Semantics emerge from Collaborative Verbosity. In *Proceedings of the 19th International World Wide Web Conference* WWW '10. Raleigh, NC, USA: ACM.

Kozima, H. (1993). *Computing Lexical Cohesion as a Tool for Text Analysis*. Technical Report.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259–284.

Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. on Knowl. and Data Eng.*, *15*, 871–882.

Manabu, O., & Takeo, H. (1994). Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th conference on Computational linguistics - Volume 2* COLING '94 (pp. 755–761). Stroudsburg, PA, USA: Association for Computational Linguistics.

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). *Introduction to Information Retrieval*. (1st ed.). New York, NY, USA: Cambridge University Press.

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *Proceedings of the 18th International Conference on World Wide Web* WWW '09 (pp. 641–650). New York, NY, USA: ACM.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *PSYCHOLOGICAL REVIEW*, *100*, 254–278.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, *38*, 39–41.

Miller, G. A., & Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, *6*, 1–28.

Milne, D. (2008). Computing Semantic Relatedness using Wikipedia Link Structure. In *Proceedings of the New Zealand Computer Science Research Student Conference* NZCSRSC '07.

Milne, D., & Witten, I. H. (2008). An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *In Proceedings of the Conference on Artificial Intelligence* AAAI '08.

Nakayama, K., Hara, T., & Nishio, S. (2008). Wikipedia Link Structure and Text Mining for Semantic Relation Extraction Towards a Huge Scale Global Web Ontology.

Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *32*, 678–692.

Navigli, R., & Ponzetto, S. P. (2012a). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217 – 250.

Navigli, R., & Ponzetto, S. P. (2012b). Babelrelate! a joint multilingual approach to computing semantic relatedness. In *AAAI Conference on Artificial Intelligence*.

Patwardhan, S. (2006). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *In: Proceedings of the EACL* (pp. 1–8).

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004* HLT-NAACL–Demonstrations '04 (pp. 38–41). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ponzetto, S. P., & Strube, M. (2007a). Deriving a Large Scale Taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2* (pp. 1440–1445). AAAI Press volume 2 of *AAAI '07*.

Ponzetto, S. P., & Strube, M. (2007b). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Int. Res.*, *30*, 181–212.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, *19*, 17–30.

Resnik, P. (1998). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, *8*, 627–633.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Salton, G., & Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, *24*, 513–523.

Schuetze, H., & Pedersen, J. O. (1997). A Cooccurrence-based Thesaurus and two Applications to Information Retrieval. *Information Processing and Management*, *33*, 307–318.

Singer, P., Niebler, T., Strohmaier, M., & Hotho, A. (2013). Computing semantic relatedness from human navigational paths on wikipedia. In *Proceedings of the 22nd international conference on World Wide Web companion* WWW '13 Companion (pp. 171–172). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *22*, 1349–1380.

Srihari, R. K., Zhang, Z., & Rao, A. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Inf. Retr.*, *2*, 245–275.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245.

Strohmaier, M., Helic, D., Benz, D., Koerner, C., & Kern, R. (2012). Evaluation of Folksonomy Induction Algorithms. *ACM Transactions on Intelligent Systems and Technology*, *3*, 74:1–74:22.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1419–1424). AAAI Press volume 2 of *AAAI '06*.

Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Mem Cognit*, *32*, 742–51.

Trattner, C., Singer, P., Helic, D., & Strohmaier, M. (2012). Exploring the Differences and Similarities between Hierarchical Decentralized Search and Human Navigation in Information Networks. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* i-KNOW '12 (pp. 14:1–14:8). New York, NY, USA: ACM.

Turdakov, D., & Velikhov, P. (2008). Semantic Relatedness Metric for Wikipedia Concepts based on Link Analysis and its Application to Word Sense Disambiguation. In S. Kuznetsov, P. Pleshachkov, B. Novikov, & D. Shaporenkov

(Eds.), *Proceedings of the SYRCODIS 2008 Colloquium on Databases and Information Systems Saint-Petersburg, Russia, May 29-30, 2008*. CEUR-WS.org volume 355 of *CEUR Workshop Proceedings*.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

West, R., & Leskovec, J. (2012). Human Wayfinding in Information Networks. In *Proceedings of the 21st International Conference on World Wide Web* WWW '12 (pp. 619–628). New York, NY, USA: ACM.

West, R., Pineau, J., & Precup, D. (2009). Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence* IJCAI '09 (pp. 1598–1603). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yang, D., & Powers, D. M. W. (2005). Measuring semantic similarity in the taxonomy of wordnet. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38* ACSC '05 (pp. 315–322). Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Yazdani, M., & Popescu-Belis, A. (2013). Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artif. Intell.*, *194*, 176–202.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., & Soroa, A. (2009). WikiWalk: Random Walks on Wikipedia for Semantic Relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* TextGraphs-4 (pp. 41–49). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zhang, Z., Gentile, A., & Ciravegna, F. (2012). Recent advances in methods of lexical semantic relatedness-a survey. *Natural Language Engineering*, *1*, 1–69.