

Evaluation of biomedical ontologies through simulated navigation

Final Report to the Austrian Marshall Plan Foundation

Daniel Lamprecht *

with

Markus Strohmaier †

Mark A. Musen ‡

Natasha F. Noy ‡

Csongor Nyulas ‡

Tania Tudorache ‡

June 8 - September 10, 2012

*Marshall Plan Stipendiat, Graz University of Technology, daniel.lamprecht@student.tugraz.at

†Graz University of Technology and Stanford University, markus.strohmaier@tugraz.at

‡Stanford University

Abstract

This report summarizes my research visit at Stanford University in summer 2012. During my stay and in ongoing work, we evaluate simulated and human navigation on biomedical Wikipedia articles.

While most visits to Wikipedia are results of direct searches, human users also navigate Wikipedia to explore certain areas of interest or to look for concepts they cannot recall the name of. By making use of some sort of background knowledge on their minds, humans are able to relate articles and guess where hyperlinks may be leading. We model this behavior with the established *decentralized search* algorithm, that allows to simulate navigation with different ontologies as background knowledge.

To compare the simulations to human click data, we conducted a user study where human navigators were given wayfinding tasks on a subset of biomedical Wikipedia articles. We show that using ontologies as background knowledge in navigation simulations exhibits the same characteristics and performance as navigation by our human test subjects. We demonstrate our findings based on four biomedical ontologies and their associated Wikipedia articles.

Keywords: Navigation, Decentralized Search, Ontology Evaluation

Contents

Contents	3
List of Figures	5
List of Tables	6
Acknowledgements	7
1 Introduction	8
1.1 General Introduction	8
1.2 Topical Introduction	8
2 Related Work	11
2.1 Navigation	11
2.2 Ontology Evaluation	12
3 Materials and Methods	14
3.1 Biomedical Ontologies	14
3.2 Wikipedia Articles	19
3.3 Decentralized Search	21
3.3.1 Single-target Search	25
3.3.2 Multiple-target Search	25
3.4 User Study	27
4 Results	28
4.1 Evaluation Approaches	28
4.2 Evaluation Metrics	28
4.3 Domain-specific Evaluation	29
4.4 Cross-domain Evaluation	29
4.5 User Study	30

5 Discussion	39
6 Conclusions	40
7 References	41

List of Figures

1	Alice’s Wikipedia Navigation Scenario	10
2	Structure of ICD-10, MeSH and SNOMED-CT	15
3	Structure of the GeneOntology	16
4	Example for an infobox template used on Wikipedia	19
5	Example for a GeneOntology infobox template used on Wikipedia	21
6	List of health conditions from WebMD	23
7	Starting portals used in navigation simulations	24
8	Domain-specific results for ICD-10 and MeSH	34
9	Cross-domain results for ICD-10, MeSH and SNOMED-CT .	35
10	Cross-domain results for the User Study	36
11	Cross-domain results for the GeneOntology	37
12	Userstudy first hops	38

List of Tables

1	Characteristics of the data sets used for our work.	18
2	Examples for clusters of Wikipedia articles	26
3	Details of the user study and the compared data sets	33

Acknowledgements

This work was generously funded by a Marshall Plan Scholarship with support from Graz University of Technology. I want to thank my advisor Dr. Markus Strohmaier at Graz University of Technology for arranging my research visit to Stanford University that lead to this work. I am also very thankful for the support from all the people involved at Stanford, particularly Division Head at the Stanford Center for Biomedical Informatics Dr. Mark A. Musen, Dr. Tania Tudorache and Dr. Natasha F. Noy.

1 Introduction

1.1 General Introduction

In the summer of 2012, I was given the exceptional possibility to work on my master's thesis in cooperation with Stanford University in California. The Stanford Biomedical Informatics Research Center has a well-established expertise in the field of biomedical ontologies. In the past, this has already led to several very successful cooperations with the research group of Dr. Markus Strohmaier at Graz University of Technology. As part of our ongoing efforts to combine the expertises of professor Strohmaier's group in Graz in the field of Knowledge Management and Social Computing with professor Musen's group, this research visit has proved very fruitful.

I spent a little over three months at Stanford University, California, where I was very warmly received by professor Mark Musen and his staff. This visit would not have been possible without the generous support by the Austrian Marshall Plan foundation.

The following report aims to summarize the work accomplished in the three-month research visit.

1.2 Topical Introduction

An ontology is "an explicit formal conceptualization of some domain of interest" [6]. Ontologies consist of concepts (such as *diseases*) and links between these concepts (such as a *classification* of the concepts). An example for an ontology is the WHO's *International Classification of diseases*, which is used by insurance companies and hospitals worldwide to report morbidity statistics. Other examples of ontologies are thesauruses or WordNet, a lexical database for the English language. In our work, we were focused on biomedical ontologies, so as to make the best use of the expertise of Stanford's Biomedical Informatics Research Center.

Evaluating ontologies poses a tough problem for the research community. Common evaluation methods include comparing to an existing gold standard, thus analyzing quality by hand [6]. Ontology users often have several ontologies covering the domain of interest at hand. Comparing ontologies in an automated manner is difficult. Different ontologies (or different ver-

sions of an ontology) can be useful for different purposes, and structural properties or lexical comparisons alone might not suffice to inform decision making. The approach of task-based ontology evaluation has proved advantageous to judge the fitness of an ontology for a specific application. In this paper, we present an automated method for evaluating the suitability of ontological structures for the purpose of guiding navigation in an information network.

In the following, we make use of a subset of biomedical Wikipedia articles to simulate human navigation. We show that several different biomedical ontologies can be used as background knowledge to inform navigation simulations, much as humans use their acquired knowledge for navigation. We demonstrate that user behavior resembles our simulations and present results of a simple user study. We suggest that our method can be used as an automated evaluation method for ontologies using these simulated navigation scenarios.

To illustrate our work, let us introduce the following example (depicted Figure 1): Alice has been diagnosed with a specific respiratory disease by her physician. Back home, she has forgotten the exact name of her condition. She decides to go and look for it on Wikipedia. Since she does not know the exact name of her target article, she cannot use the search function. Alice starts from a hypothetical Wikipedia portal containing links to a number of common diseases. She first clicks the portal link leading to the article on *Asthma*, as this seems to be a good starting point for finding her respiratory disease. Next, she navigates to *Chronic Bronchitis*, then to *Pneumonia* and finally arrives at *Bronchopneumonia*, which she recognizes as the disease the doctor diagnosed her with.

At each step in her navigation, Alice only sees the links pointing away from the article she currently is at. She is familiar with some of the articles, and is able to relate them to one another through what we refer to as her *background knowledge*. She recognizes some of the hyperlinks and knows what their target article will likely be about. Since Alice is only making use of the *local* article content and its outgoing links at each step, she performs what is called *decentralized search* in the literature [11].

In our work, we demonstrate the use of biomedical ontologies as background knowledge for navigation. We simulate navigation by means of decentralized search on a set of biomedical Wikipedia articles. We show that several biomedical ontologies show the same characteristics and performance as actual human test subjects.

Our main contributions are the demonstration of the general suitability of

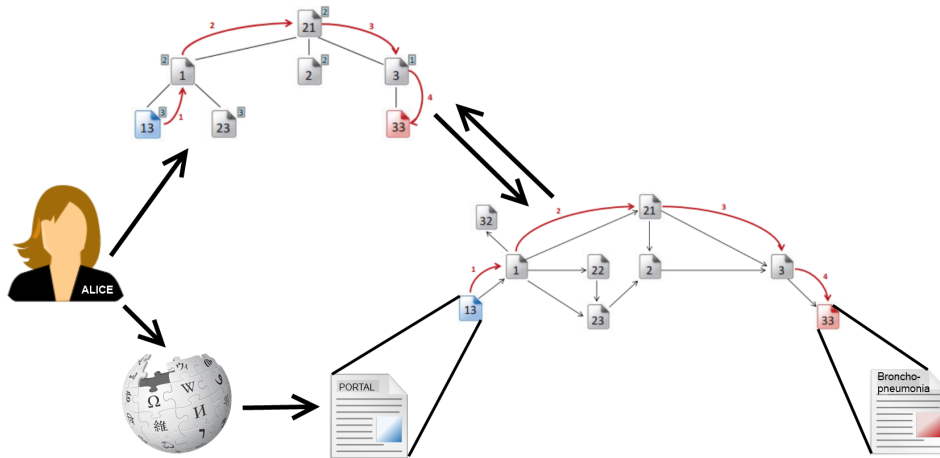


Figure 1: **Alice’s Wikipedia Navigation Scenario** Looking for a disease, Alice goes to Wikipedia and starts from a hypothetical portal containing links to a number of common diseases. Alice then navigates her way through the Wikipedia network. Since she does not know the exact name of her target, she does not use the search function. Being familiar with some of the articles, she is able to relate them to one another through her background knowledge. Guessing in what area hyperlinks are leading and how likely they will lead her in the right direction, she is able to find the disease she is looking for.

external real-world ontologies to inform decentralized search on Wikipedia and their comparison on the same data set as well as with upper and lower bounds. Our findings are relevant for researchers interested in new evaluation methods for ontologies and researchers interested in modeling network navigation.

The rest of this paper is structured as follows: In section 2 we place our work in the context of previous research and related work. In Section 3 we discuss materials and methods, and we present our results in Section 4. We end with a discussion of our results.

2 Related Work

The research related to this work can be broadly divided into *navigation* and *ontology evaluation*.

2.1 Navigation

Decentralized search, as used in our navigation simulations, was made famous by Stanley Milgram’s widely discussed small-world experiment [17] in the 1960s. In the experiment, participants in Boston and Nebraska received a letter containing information about a target person (a Boston stock broker). They were then asked to forward the letter to one of their acquaintances so as to bring the letter closer to the target person. The forwarding was limited to close friends, i.e., acquaintances with whom participants corresponded on a first-name basis. Each participant was further asked to record information in the letter, such as profession and age.

The results showed a length of four to six hops for successful chains of letters. The letter distribution took a bimodal form, stemming from the two main ways in which letters reached the target person: Through his work and through his hometown. By taking only the limited knowledge of each participant into account, the search effectively constituted a form of decentralized search. The result demonstrated the *small world phenomenon*, as it seemed possible to connect two arbitrary persons across the United States through a very small number of hops. It was later argued that this theory would hold for the whole world in general. In 2011, researchers at Facebook showed that the social network’s users were a mere four hops away from each other [4].

Some thirty years later, the theory of small world models was extended by Watts and Strogatz [18] who showed some of the characteristics of graphs displaying the small world phenomenon. They also demonstrated the occurrence of small world networks in a power grid, a neural network of a worm and an actor collaboration network.

In 2000, Jon Kleinberg showed that no decentralized algorithm could exist for the type of small world network proposed by Watts and Strogatz [11]. However, Kleinberg presented a more generalized version of the model, for which he proved that a decentralized algorithm capable of finding short paths existed.

As far as navigation on Wikipedia is concerned, recent research [7] has shown that when visiting a Wikipedia page, users have a 30 - 40% chance of following a link on that page. Users are hence more likely to jump to some other page directly. Jumping to another page is referred to as *teleporting*, e.g. by using the search function or typing in another address manually. In general, users are estimated to follow a hyperlink in about 60 - 70% of their clicks. The fraction of teleports is hence significantly higher on Wikipedia than on general web sites. This might be due to the fact that users visit Wikipedia to satisfy specific information demands rather than browsing the articles. However, navigating Wikipedia does occur, e.g., in the examples given in Section 1.

Navigating Wikipedia has been studied in the context of *Wikigames*. West et al. [20] used Wikigame data to infer semantic distances between concepts by studying game click paths. West and Leskovec [19] found that in Wikigames, players tend to navigate to hubs (articles with a large number of outlinks) first and subsequently home in on targets node.

Previous research in our own group has focused on simulating decentralized search in navigation networks. In [9], the authors evaluated different folksonomies using decentralized search. In [15], Strohmaier et. al. compared different folksonomy induction algorithms through decentralized search.

2.2 Ontology Evaluation

In 2005, Brank et al. [6] surveyed ontology evaluation techniques and identified four main methods:

1. *Comparison to a gold standard* by lexical comparison to an ontology considered a good representation, and measuring deviation from that ontology.
2. *Data-driven evaluation* by measuring the fit of an ontology to data, e.g., a set of documents - e.g., measuring the overlap of key terms in the ontology and the document set. Other measures for this include ontology breadth and depth [22], which measure domain coverage.
3. *Manual assessment* by humans, who analyze the ontology with regard to requirements and standards. While this is clearly the best evaluation method, it is often not feasible. Particularly for large ontologies or for evaluation of several revisions of an ontology, it is too time-

consuming or expensive to concern human experts with evaluation tasks.

4. *Task-based evaluation* by measuring the fitness of an ontology for a given task or application. This establishes the suitability for one specific task only - but this is often sufficient for real-world applications.

The task of automatically evaluating ontologies is still considered hard in the research community.

For this paper, we focused on *task-based* evaluation by using simulated navigation and compared different ontologies as background knowledge. Previous research [16] has already analyzed some of the aspects of decentralized search and human search behavior on Wikipedia. In their work, Trattner et al. showed that background knowledge based on structural network features performed better than external background knowledge. In this paper, we again make use of the simulated decentralized search to study ontological structures. We compare different external ontologies as background knowledge for decentralized search and analyze their differences as well as similarities with human navigation behavior.

3 Materials and Methods

Our data consisted of four biomedical ontologies and corresponding articles from the English Wikipedia. We simulated decentralized search on the subset of the Wikipedia link network induced by these articles. The ontologies served as the background knowledge. To validate our results, we conducted a user study and compared the results to our simulations.

3.1 Biomedical Ontologies

We used the following four ontologies from the biomedical domain for our research:

The **International Classification of Diseases, tenth revision (ICD-10)** is a classification of diseases, signs and symptoms first published in 1992 and maintained by the World Health Organization (WHO). ICD-10 had its origins in the classification of causes of deaths and is presently used by over 100 countries to report mortality statistics. It is also widely used for epidemiology, health management as well as clinical purposes and is available in 46 languages [1]. The version we used contained 12,417 concepts. ICD-10 consists of 22 top-level nodes termed *chapters* and assigns a code (or a range of codes) to every disease in its domain. ICD is currently being developed in its 11th revision. For this process, the WHO is using collaborative ontology engineering tools for the first time in the history of ICD. This process was the subject of another research cooperation with Stanford University and yielded several publications, to one of which this author contributed [10].

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus for journal articles in the medical domain. MeSH is maintained by the U.S. National Library of Medicine. The ontology forms a tree-structure with 16 top-level concepts and contains 26,142 terms (dubbed *descriptors*) [2]. Descriptors are graph leaves and attached to one or more tree nodes (which are not descriptors). As such, the complete ontology graph we used actually contained 80,689 nodes. MeSH extends beyond biomedical concepts and comprises terms from other domains such as Geography, Technology or Publication Characteristics.

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) [14] is a clinical healthcare terminology used in electronic health record

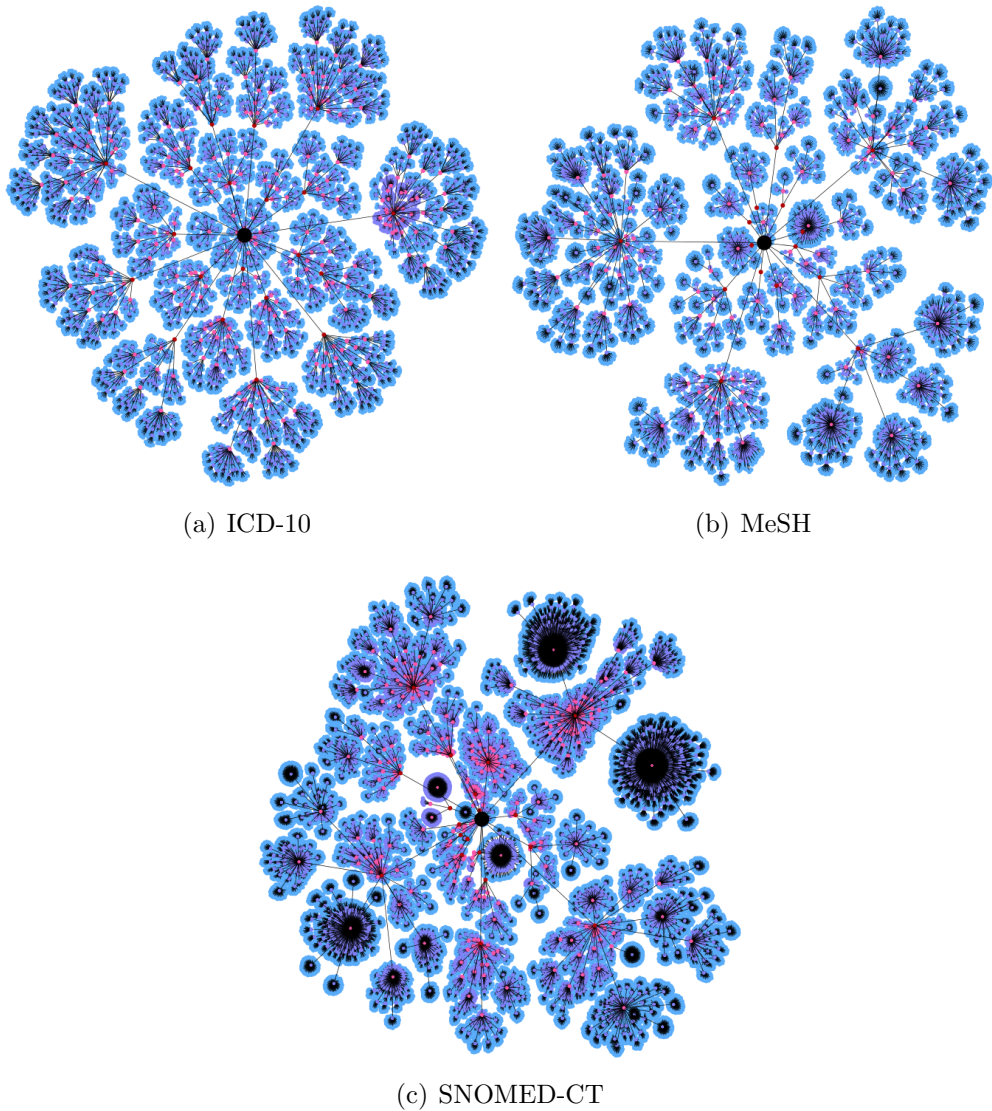


Figure 2: **Structure of the four top levels of ICD-10, MeSH and SNOMED-CT.** The figure shows the structure for ICD-10 and MeSH, followed by SNOMED-CT. The root node is displayed in the middle of each plot. The figures show all ontology concepts up until a distance of four from the root node (hence the term "levels"). Color indicates distance, with red being close to the root and blue being farther away. SNOMED-CT (depth 16) is clearly broader than MeSH (depth 14), which stems from the fact that the latter contains roughly four times as many concepts as the former.

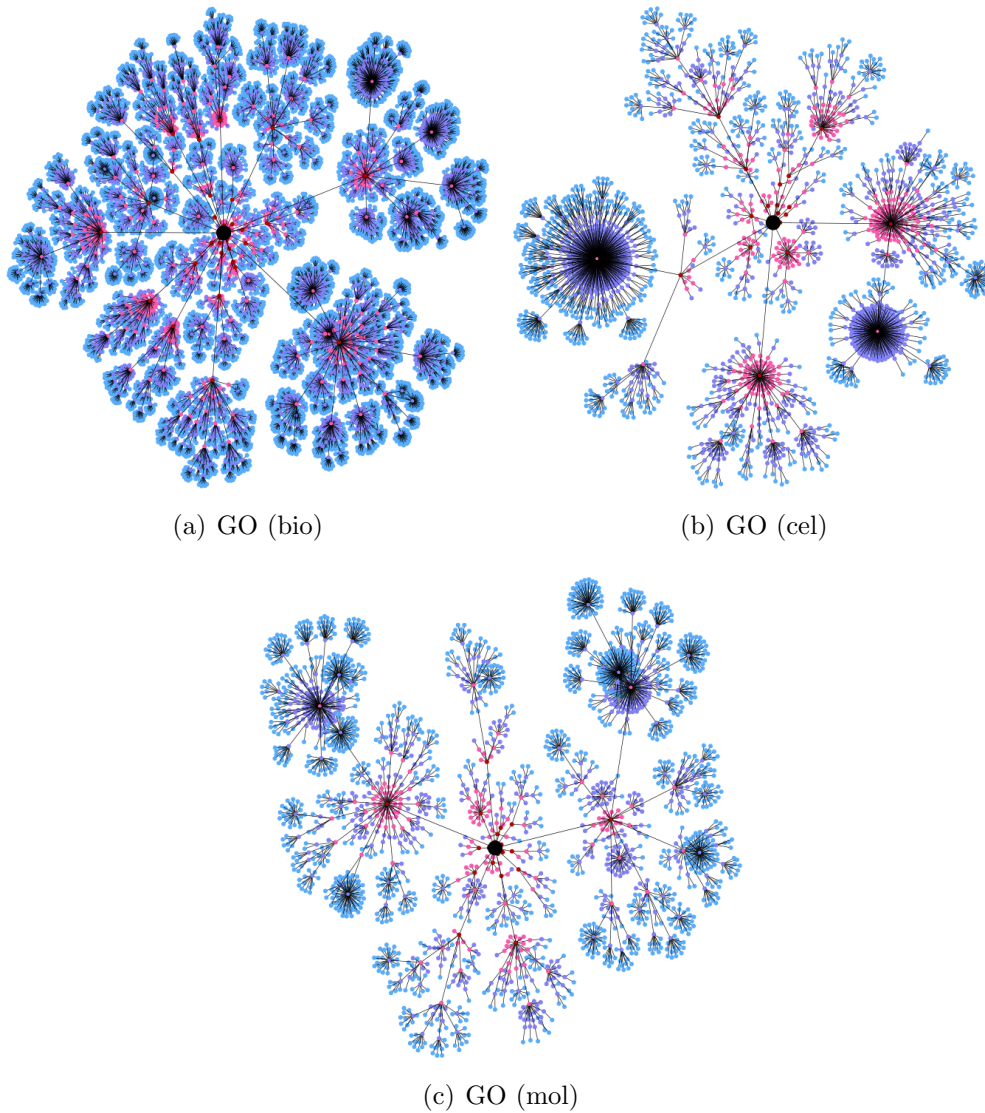


Figure 3: **Structure of the four top levels of the GeneOntology.**

The figure shows the structure for the three subontologies making up the GeneOntology (biological process, cellular component and molecular function). The root node is displayed in the middle of each plot. The figures show all ontology concepts up until a distance of four from the root node (hence the term "levels"). Color indicates distance, with red being close to the root and blue being farther away.

systems. The revision we used contained 295,482 concepts, which made it by far the largest ontology in our simulations. SNOMED-CT consists of 19 top-level concepts. In contrast to the other ontologies, SNOMED-CT is proprietary.

The **Gene Ontology (GO)** [3] is a controlled vocabulary of terms used for the annotation of human genes and gene products. It consists of 37,779 concepts divided among three different subontologies, which cover the cellular component, the molecular function and the biological process, respectively. In its filtered form which we used for our study, the three subontologies take the form of disjoint trees.

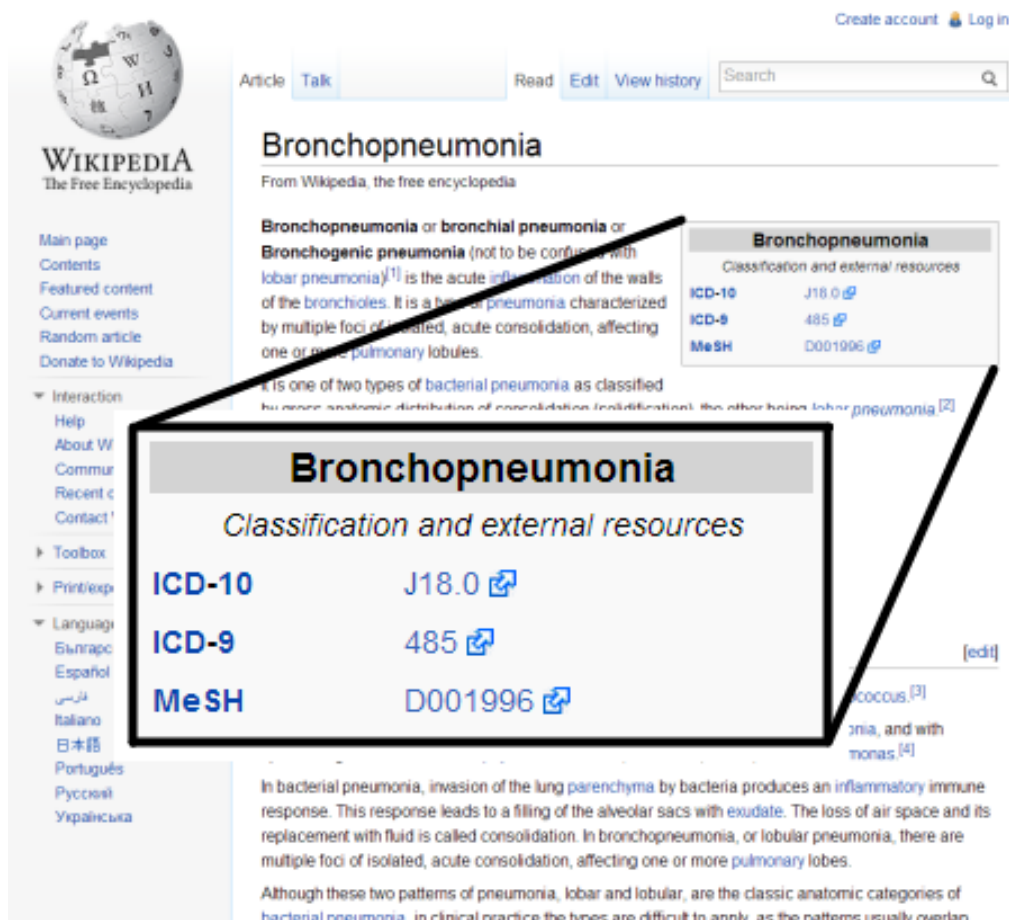
Table 1 displays statistics about the data sets used for this paper. Figures 2 and 3 depicts the first five levels of the used ontologies.

	Description	ICD-10	MeSH	SNOMED-CT	Gene Ontology	bio	cel	mol
Ontology	concepts	12,417	80,689	295,482	37,779	23,691	3,020	9,413
	top-level links	22	16	19		25	19	21
	density	12,416	112,463	440,408	67,991	51,397	5,617	10,977
	depth	8.05×10^{-5}	1.73×10^{-5}	5.04×10^{-6}	4.76×10^{-5}	9.16×10^{-5}	6.16×10^{-4}	1.24×10^{-4}
Wikipedia	relation	4	14	16		10	7	10
	articles	is-a	is-a, part-of	is-a	is-a, part-of, regulates			
	links	2,673	2,584	1,594	3,445			
	density	4.46×10^{-3}	4.10×10^{-3}	5.73×10^{-3}	1.32×10^{-3}			

Table 1: **Characteristics of the data sets used for our work.** The table shows statistics about the used ontologies as well as the sets of Wikipedia articles mapped to those articles. As SNOMED-CT was only used in conjunction with MeSH and SNOMED-CT, the column shows the information for the intersection of Wikipedia articles mapping to all three ontologies. We used data from the four ontologies listed, all of which were in the biomedical domain. For the GeneOntology, the triples (bio - cel - mol) list statistics for the three subontologies making up the GeneOntology.

3.2 Wikipedia Articles

We used the December 2011 dump ¹ of the English Wikipedia to extract articles from the biomedical domain corresponding to ontology concepts. The dump is made available in XML format, which we preprocessed and filtered for relevant articles. We then mapped the articles to the ontologies by parsing the articles' info boxes, i.e. the template structures containing the information linking to biomedical ontologies.



The image shows a screenshot of a Wikipedia article titled "Bronchopneumonia". The article content is partially visible, describing the condition as an acute inflammation of the bronchioles. A callout box highlights the "Bronchopneumonia" infobox template, which is titled "Classification and external resources". This template contains the following information:

Bronchopneumonia	
Classification and external resources	
ICD-10	J18.0 [e]
ICD-9	485 [e]
MeSH	D001996 [e]

Below the infobox, the article text begins with "In bacterial pneumonia, invasion of the lung parenchyma by bacteria produces an inflammatory immune response. This response leads to a filling of the alveolar sacs with exudate. The loss of air space and its replacement with fluid is called consolidation. In bronchopneumonia, or lobular pneumonia, there are multiple foci of isolated, acute consolidation, affecting one or more pulmonary lobes. Although these two patterns of pneumonia, lobar and lobular, are the classic anatomic categories of bacterial pneumonia, in clinical practice the types are difficult to apply, as the patterns usually overlap."

Figure 4: **Example for an infobox template used in disease articles on Wikipedia.** Disease articles commonly make use of an Infobox disease template, which offers fields for ontology codes. We used template fields in the Infoboxes to map Wikipedia articles to their ontology counterparts.

Disease articles commonly make use of a `Template:Infobox disease`²,

¹<http://dumps.wikimedia.org/enwiki/20111201/>

²http://en.wikipedia.org/wiki/Template:Infobox_disease

which offers several options to reference medical ontologies such as ICD-10 or MeSH (see Figure 4 for an example). We used template fields in the **Infobox disease** as well as a in a handful of other infobox templates to map Wikipedia articles to their ontology counterparts in ICD-10 and MeSH.

SNOMED-CT is proprietary and not present in Wikipedia info boxes. As a consequence, we could not directly relate Wikipedia articles to the ontology concepts. We therefore used semantic mappings from BioPortal [21] to map Wikipedia articles to SNOMED-CT. We mapped a total of 1,594 Wikipedia articles from both ICD-10 and MeSH to SNOMED-CT with this method. We then compared the performance of all three ontologies on this intersection of data sets.

The Gene Ontology is different in that it is not used for 1:1 mappings but for *annotation* of Wikipedia articles. Articles are assigned different annotations from the controlled vocabulary that constitutes the GeneOntology. E.g., *Insulin* is annotated with *protease binding*, *hormon activity* and *protein binding*, stemming from the *Molecular function* part of the GeneOntology.

To extract Wikipedia articles based on these templates, we went through the dumped Wikipedia articles and filtered those containing relevant templates providing ontology information.

As a result, we related Wikipedia articles to all (up until 50 or more) related concepts from all three subontologies of the Gene Ontology . We used the corresponding fields in templates created by the **ProteinBoxBot**³ to extract the relevant mappings to the Gene Ontology. The **Portal: Gene Wiki** on Wikipedia contains around 10,000 articles on human genes and proteins. Articles in this domain are usually either created or annotated by the ProteinBoxBot using information from the Gene Ontology and other projects. As a great number of these articles are very domain-specific and only very few editors are knowledgeable enough to add to them, there is a large number of stubs (very short articles) and orphans (articles not linked to by any other Wikipedia article). This is also reflected in the low number of links between the articles in this data set, as compared to the other data sets (see the density information in Table 1).

³<http://en.wikipedia.org/wiki/User:ProteinBoxBot>

Solute carrier family 26 (sulfate transporter), member 2	
Identifiers	
Symbols	SLC26A2; D5S1708; DTD; DTDST; EDM4; MST153; MSTP157
External IDs	OMIM: 606718 MGI: 892977 HomoloGene: 73876 GeneCards: SLC26A2 Gene
Gene Ontology [hide]	
Molecular function	<ul style="list-style-type: none"> • secondary active sulfate transmembrane transporter activity • sulfate transmembrane transporter activity
Cellular component	<ul style="list-style-type: none"> • plasma membrane • integral to plasma membrane • membrane
Biological process	<ul style="list-style-type: none"> • ossification • ion transport • sulfate transport • transmembrane transport
Sources: Amigo / QuickGO	

Figure 5: **Example for a GeneOntology infobox template used on Wikipedia.** Articles are annotated with multiple elements taken from the three subontologies of the Gene Ontology. We used those annotations to relate Wikipedia articles to GeneOntology concepts.

3.3 Decentralized Search

In this section, we'll first go over the general algorithm for decentralized search, followed by a detailed description of its application to our research.

In general, decentralized search works as follows: A start node of a network (e.g., a person in a social network) is provided with information about the target node (e.g., another person). The start node's task is then to forward the search task to one of its adjacent nodes (e.g., friends), so as to forward it towards the target node as quickly as possible. The "decentralized" part about it is, that at no point, global knowledge is available. Each node along the search path simply uses its local neighborhood knowledge

to forward the search problem.

A real world example for this type of search could be fixing a bug in a piece of software: When a user finds a bug, they enter it in the bug fix system. The developer in charge of the bug fix system then assigns the bug to the developer they consider most likely to fix the bug. When we assume that the software development team consists of a large number of people, the first person might not necessarily be the right one to deal with the issue. However, this person might guess who could be responsible for the specific section of the software, and reassign the bug. If we imagine that this process is repeated several times before the right developer is found, this is exactly decentralized search as described before.

To simulate Alice’s usage of Wikipedia, we mapped all articles to their corresponding ontology counterparts. Given these mappings, the simulation could then calculate distance information on the ontology (the *background knowledge*). The distance from a start to a target article in the ontology was taken as the length of the shortest path that connected these concepts not in the Wikipedia graph but in the ontology. In this manner, the simulator had the full knowledge about the ontology at its disposal, just as Alice could access her whole background knowledge in her mind. The simulator would then use this information to determine the best link to click, i.e., the link for which the ontology predicted it would lead to the article closest to the target article. The distance information gained this way was not necessarily optimal or even correct, but provided a good guess to guide navigation.

In our simulation, the target article was directly known to the simulation. This was used to model the somewhat familiar article Alice was trying to reach. Alice did not know the exact name of her target, but she could roughly place it in a category, to which she then navigated using her own background knowledge. Our simulations modeled this by calculating distance directly to the target node on the background knowledge to determine the best link to click.

To avoid loops, the simulation visited each link in the network only once (but could visit nodes multiple times). In some cases a node yielded only links which, according to the ontology distance information, would put the simulation in a worse state than before, or no (new) links at all (i.e., a dead end). In these cases the simulation would backtrack to the last visited node, just as Alice would use her browser’s *back button*. At any given point, the simulation could also jump back to the starting portal directly, modeling a *home button*.

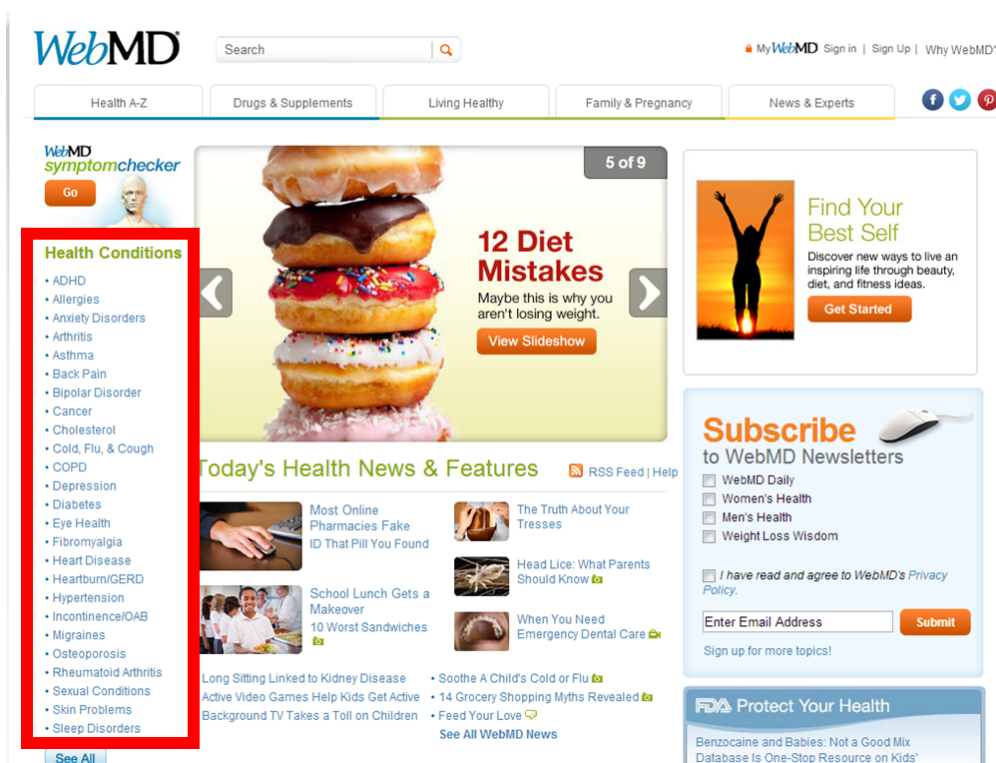


Figure 6: List of health conditions from WebMD We used the list of health conditions from WebMD.com as our hypothetical Wikipedia starting portal for our disease articles. We manually mapped these conditions to Wikipedia articles in our data set and started the simulations from the page containing links to them. We WebMD to best reflect our search scenario of users looking for somewhat familiar diseases.

We started the navigation from a hypothetical Wikipedia portal featuring a selection of suitable articles. For ICD-10, MeSH and SNOMED-CT, we used the 25 health conditions listed in the navigation bar of WebMD.com (see Figures 6 and 7). We manually mapped these conditions to Wikipedia articles from our dataset and used the articles as the outgoing links from the portal. We chose WebMD to best reflect our search scenario of users looking for somewhat familiar diseases. For the Gene Ontology, we used the articles listed in the two top-10 lists (ranked by word count and view count) shown on the Portal: Gene Wiki⁴ (see table 7).

In our simulations we analyzed two different search scenarios, which we describe in the following two subsections.

⁴http://en.wikipedia.org/wiki/Portal:Gene_Wiki

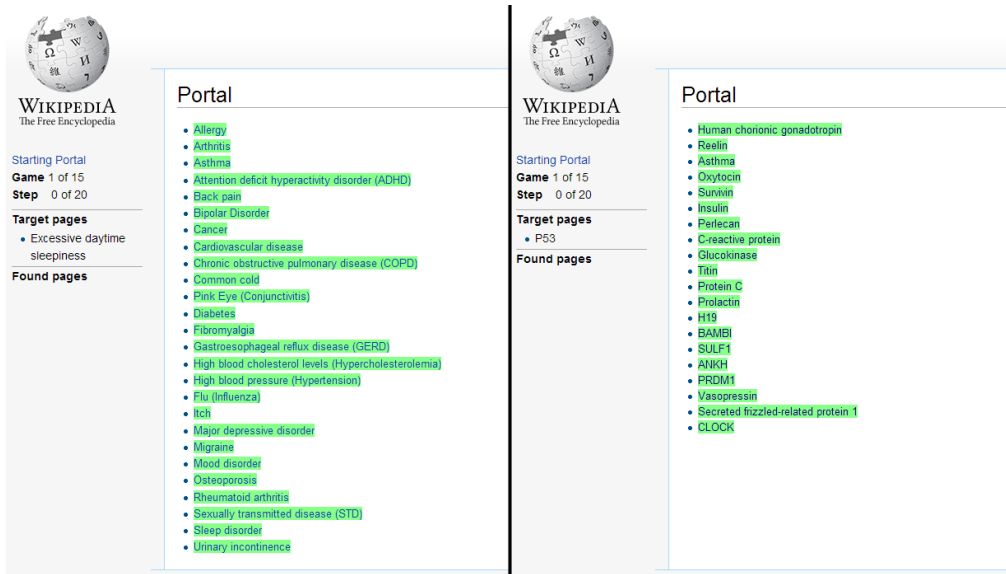


Figure 7: **Starting portals used in navigation simulations.** For ICD-10, MeSH and SNOMED-CT we used a portal obtained by mapping navigation bar articles from WebMD.com to Wikipedia articles. For the GeneOntology, we used the longest and most frequently visited Wikipedia articles as listed on the Gene Wiki Portal http://en.wikipedia.org/wiki/Portal:Gene_Wiki.

3.3.1 Single-target Search

Our first scenario was analogous to Alice’s example of looking for a disease. In single-target search, the simulation started at the hypothetical Wikipedia portal as described and proceeded to a single target article using decentralized search.

As discussed, single-target search modeled the scenario of having something on the tip of one’s tongue, and navigating to rediscover it.

3.3.2 Multiple-target Search

For multiple-target search we used the same approach as for single-target search. The only difference was in the targets, which consisted of target sets of 2 through 10 articles. The rest of the simulation (starting portal, decentralized search, background knowledge) was conducted in the same way as the single-target search.

We used multiple-target search to model a scenario of exploratory search. In exploratory search, users explore a space of resources rather than trying to find one specific target. We measured the successfulness of the exploration by the number of found targets.

We used clusters of semantically similar Wikipedia articles as our target sets. To automatically obtain these clusters, we first extracted textual features from the articles. The type of features we used were TF-IDF (Text Frequency - Inverse Document Frequency). To be more specific, this means that each article is represented as a vector of all the words occurring in all articles, and a weight for each entry in this vector. The weight is determined by the occurrence of the specific word in the article and its occurrence in all other articles. The idea is, that words occurring only in one article are highly representative for that article, but words that also occur in every other article are not. After the feature extraction, we then used *k-means clustering* to arrange similar articles into clusters. We used those resulting clusters containing two through fifteen articles as target clusters in our simulations. Examples for clusters are given in Table 2. We used the Scikit-learn library [13] for this part of our work.

Nausea-related	Stomach-related	Cough-related
Vomiting Nausea Motion sickness Morning sickness Drooling Hyper. gravidarum	Linitis plastica Stomach cancer Gastritis Atrophic gastritis Ménétrier’s disease Achlorhydria Gastroparesis Duodenal cancer Gastric dumping syndrome Stomach disease	Bronchitis Chronic bronchitis Acute bronchitis Cough Sputum
Cancer-related	Personality disorders	
Stomach cancer Breast cancer Pancreatic cancer Prostate cancer Cancer Lung cancer Leukemia	Avoidant personality disorder Borderline personality disorder Antisocial personality disorder Panic attack Schizoid personality disorder Obsessive–compulsive disorder	

Table 2: **Examples for clusters of Wikipedia articles used in exploratory search.** The table shows three examples of clusters used in our simulations. We used TF-IDF features and k-means clustering to automatically group Wikipedia articles into semantically related groups of two through ten articles.

3.4 User Study

To evaluate our simulations, we carried out a user study on Wikipedia navigation. Eight participants without any particular background in medicine were asked to navigate Wikipedia, modeling the scenario of navigating to find diseases. The study used the intersection data set of ICD-10, SNOMED-CT and MeSH, containing 1,594 Wikipedia articles. As a large share of these articles turned out to be too specialized for test subjects not particularly familiar with the medical domain (with article names such as *Halitosis*, *Aniseikonia* or *Milroy's disease*), we manually selected 100 generally better known targets (such as *Pneumonia*, *Stomach cancer* or *Asthma*), out of which we also manually formed 20 clusters of four articles each. We then set up the subset of Wikipedia articles in our testing environment and asked subjects to perform navigation tasks. As in our simulations, backtracking (using the back button in the browser) and jumping back to the portal by clicking a link were enabled at all times.

Our simulation tasks were analogous to what is commonly called *Wiki Games*. Wiki Games, such as <http://wikispeedia.net/> or <http://thewikigame.com/> provide the user with a start and a target page. The aim of these games is to navigate to the target page without using the search function or domain-external pages, i.e. by only clicking links in the article body text. Game success and ranking is then established by different criteria, e.g. minimum number of clicks or minimal time. Over time, multiple variants of Wiki Games have been developed, such as *Five Clicks To Jesus* ⁵, which requires the users to navigate to the article on Jesus in a maximum of five clicks.

Each participant completed a total of 15 games/navigation tasks. As a starting point, we used the hypothetical WebMD Wikipedia portal also used by our simulator (see Figure 7). To deal with potential frustration, participants were given the possibility to abort the current task if they had not found the target(s) after half of the maximum number of steps (20 for single targets and 40 for multiple targets).

⁵<http://thewikigame.com/5-clicks-to-jesus>

4 Results

4.1 Evaluation Approaches

We used the following evaluation approaches:

1. Firstly, for the **domain-specific** evaluation we mapped each ontology to the maximum number of articles available. We then evaluated the performance for each ontology on its *domain-specific set of articles* (i.e., each ontology on a different set of articles). This allowed us to evaluate the fitness of each ontology for its specific domain of interest.
2. Secondly, we evaluated the **cross-domain** performance of several ontologies. For this, we reduced the set of Wikipedia articles to the intersection, i.e., the *set of articles mapping to all examined ontologies*. We could then directly compare the performance of different ontologies on the same data set. This allowed us to compare ontology behavior directly.
3. Thirdly we juxtaposed three ontologies to *human behavior* by comparing to a **user study**. We picked a set of tasks for navigation tasks and asked our test subjects to navigate to them. We then ran a simulation with the same data set and targets and evaluated the results. This allowed us to compare our simulations to human behavior.

4.2 Evaluation Metrics

Based on [12] and [5] we used *stretch* and *success ratio* to evaluate navigation paths.

We define *success ratio* s to be the average fraction of target nodes found and *stretch* τ to be the average ratio of found path lengths to shortest path lengths. As in [8] and [16] we evaluate success ratio and stretch broken down by shortest path length of the underlying node pairs. These metrics give us a means of analyzing what paths were found by the simulator and how much longer than the shortest paths they were.

We further extend these metrics with the *accumulated success ratio* as , which we define as the average fraction of nodes found until a certain

number of steps.

For all our evaluations, we assumed a maximum number of 20 clicks for single targets and 40 clicks for multiple targets.

4.3 Domain-specific Evaluation

For the domain-specific evaluation, we compared ICD-10 to MeSH. We used two different sets of Wikipedia articles, namely the whole sets of 2,673 articles mapping to ICD-10 and the 2,584 articles mapping to MeSH.

Figure 8 shows the results. All four metrics indicate that ICD-10 performed better for targets closer to the portal, and MeSH better for targets farther away. The Success Ratio shows that while MeSH found slightly more paths overall, ICD-10 found more paths with an underlying shortest path length of two. This is also reflected in the Accumulated Success Ratio. The Stretch shows that the paths found by ICD-10 were, on average, slightly shorter than the ones found by MeSH.

4.4 Cross-domain Evaluation

For the cross-domain evaluation, we evaluated multiple ontologies on the same set of Wikipedia articles. Cross-domain evaluation allowed us to inspect multiple ontologies side by side, facilitating comparison.

The data sets we used for this were (i) the set of the articles mapping to both ICD-10 and MeSH as well as to SNOMED-CT and (ii) the set of articles mapping to all three subontologies of the GeneOntology.

For (i) Figure 9 depicts the results. We established upper and lower bounds by including both a random walk and an optimal solution. We averaged over 1000 random walks for each target node to obtain the lower bound. The optimal solution was calculated using global knowledge of the Wikipedia network. For the single-target search, the optimal solution was obtained by using global knowledge (and hence the globally shortest paths). For the multiple-target search, a nearest-neighbor approach with global graph knowledge was used to approximate an optimal solution because a truly optimal solution would require solving an instance of the Traveling-Salesman problem, which is computationally expensive. The results show that the simulations were, for the most parts, well between the

optimal and the random solution.

The figures show that ICD-10 and MeSH performed better than SNOMED-CT for the number of paths found for both single-target and multiple-target search. In contrast to the domain-specific evaluation, ICD-10 performed better than MeSH for the Success Ratio and the Accumulated Success Ratio (although they were still close to each other). This might be due to the intersection of articles removing some targets which were harder to find. In terms of Stretch, ICD-10 found shorter paths than SNOMED-CT, which in turn fared slightly better than MeSH. This is consistent with the domain-specific evaluation, where ICD-10 found shorter paths than MeSH. The stretch for the random walk is in the range of the ontology results for shortest paths of length four. However, this is due to the random walk finding only a tiny fraction of the targets for a distance of four, thus biasing the resulting stretch.

Figure 11 shows the results for (ii). The GeneOntology data set had noticeably worse results than the other data sets considered in this paper. With target nodes as far away as eleven hops from the starting portal, the ontology-informed search was only able to find targets up to a distance of six hops and no more than 25% of the targets at a distance of two hops. Overall, the ontologies were only able to find between four and seven percent of target nodes.

4.5 User Study

For the user study, we compared the performance of human navigators with the ontologies on the same data set used for the *cross-domain* evaluation of ICD-10, MeSH and SNOMED-CT. The targets were 100 manually selected targets and 20 manually selected clusters. To deal with potential user frustration when not finding targets, we allowed aborting the search after half the number of clicks. In our analysis, we assumed aborted tasks to contain the maximum number of clicks. The limitation of targets also meant that targets were a maximum distance of three hops away from the portal. The evaluations do hence not include any data points for longer shortest paths.

Figure 10 shows that the success ratio for the user study was fairly close to the simulator performance. The overall success ratio was 92% for the user study and ranged from 79 - 91% for the ontologies. That is, human test subjects were able to find slightly more target articles than the simulation in single-target search. For multiple-target search, the accumulated success

ratio shows that the user study fell within or just below the range of the three ontologies. Noticeable is that after 20 steps, the user study did not find any more targets. This coincides with the point from where on users were allowed to abort their search task. For the single-target stretch again, with an overall stretch of 1.74 the user study performed slightly better than the ontologies, which displayed stretches between 1.78 and 1.84.

In addition, we compared several further aspects of the user study to the ontologies. Table 3 displays these statistics and compares the user study to the ontologies as well as the optimal solution and the random walk.

In summary, the results confirm that what has appeared somewhat apparent from the Success Ratios and the Stretch, i.e., that ICD-10 and MeSH displayed the most similar behavior to the user study.

To compare the values, we calculated cosine similarity. Cosine similarity works by comparing vectors of values. In our case, we arranged our data into vectors (e.g., a vector of all targets and a "1" for "found" and a "0" for "not found"). Cosine similarity then calculates the angle between two vectors as

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=0}^N A_i \times B_i}{\sqrt{\sum_{i=0}^N (A_i)^2} \times \sqrt{\sum_{i=0}^N (B_i)^2}}$$

For vectors with nonnegative elements this always yields a value between 0 and 1 (where one denotes vector equality and 0 orthogonality).

For the found targets, all three ontologies displayed high cosine similarity values. This is due to the high success ratios for the limited target set used in our user study which leads to the majority of the vectors containing ones at the same positions. The random walk (which found only (4%) of the targets) also sets a high base line of 72%.

The visited Wikipedia pages showed greater differences. Interestingly enough, the random walk was about as similar or more similar to the user study as the three ontologies for the single-target search. The optimal solution was even more similar. This is due to the large number of graph nodes (around 90%) that are never visited in this study but are still included in the calculation. Since nodes visited by the optimal solution lie on the shortest path to the target, the simulation is likely to visit some of these nodes as well. For the ontologies, MeSH displayed the most similar behavior to the user study.

For the first hops (i.e., the first click on a portal link), the clicks were distributed quite evenly. A random distribution would see each link clicked 3.7% of times. Our results showed distributions ranging from 1 through 17% and were thus fairly evenly distributed, explaining the values of the cosine similarity being close together. For the first hops, ICD-10 displayed the most similar values to the user study. Figure 12 shows the first hop distribution in more detail.

In addition to calculating similarities, we also inspected the average per-step probability of backtracking or clicking the home button.

Both the simulation and the users had access to a back button (leading to the previously visited page) and a home button (leading back to the portal) at all times. The simulations used the home button only immediately after having found a target in multiple-target search. In all other cases, the best strategy given by our simulation constraints turned out to be backtracking or not even visiting the node in the first place. The user study showed different behavior from the simulator in several aspects: For single-target search, users backtracked less frequently (9% of clicks were back button clicks, versus 11-13% for the simulations) but used the home button in 2% of clicks. For the multiple-target search, users backtracked more frequently (27% versus 17-18% for the simulator) and used the home button less frequently (1% versus 2-3%).

In conclusion, backtracking was the most widely applied strategy for navigating out of dead ends and backtrack from less promising areas of the network. This was especially true for the user study.

	User Study	ICD-10	MeSH	SNOMED-CT	Optimal	Random
Found targets (Cosine Similarity)	Single	0.93	0.95	0.89	0.95	0.72
	Multiple	0.94	0.94	0.91	0.95	0.67
Visited Wikipedia Pages (Cosine Similarity)	Single	0.60	0.71	0.58	0.87	0.69
	Multiple	0.51	0.75	0.37	0.68	0.39
First Hops (Cosine Similarity)	Single	0.89	0.85	0.69	0.88	0.80
	Multiple	0.64	0.62	0.56	0.68	0.71
Back Button Uses (average per step)	Single	0.13	0.11	0.13	0.00	0.07
	Multiple	0.17	0.18	0.18	0.01	0.09
Home Button Uses (average per step)	Single	0.00	0.00	0.00	0.00	0.00
	Multiple	0.03	0.02	0.03	0.00	0.00

Table 3: **Details of the user study and the compared data sets** The table displays statistics about the user study and the ontologies. The most similar values to the user study are displayed in bold face. For the first three measures, we viewed the information about found targets, visited Wikipedia Pages and first hops as a vector of values, for which we calculated the angle to the vector containing the information for the user study (i.e., the cosine similarity). For the random walk, we averaged over 1000 random walks for each portal-target pair. The last two measures display the average per step usage of the back and home buttons for the different scenarios. In summary, the results confirm that what has appeared somewhat apparent from the Success Ratios and the Stretch, i.e., that ICD-10 and MeSH displayed the most similar behavior to the user study.

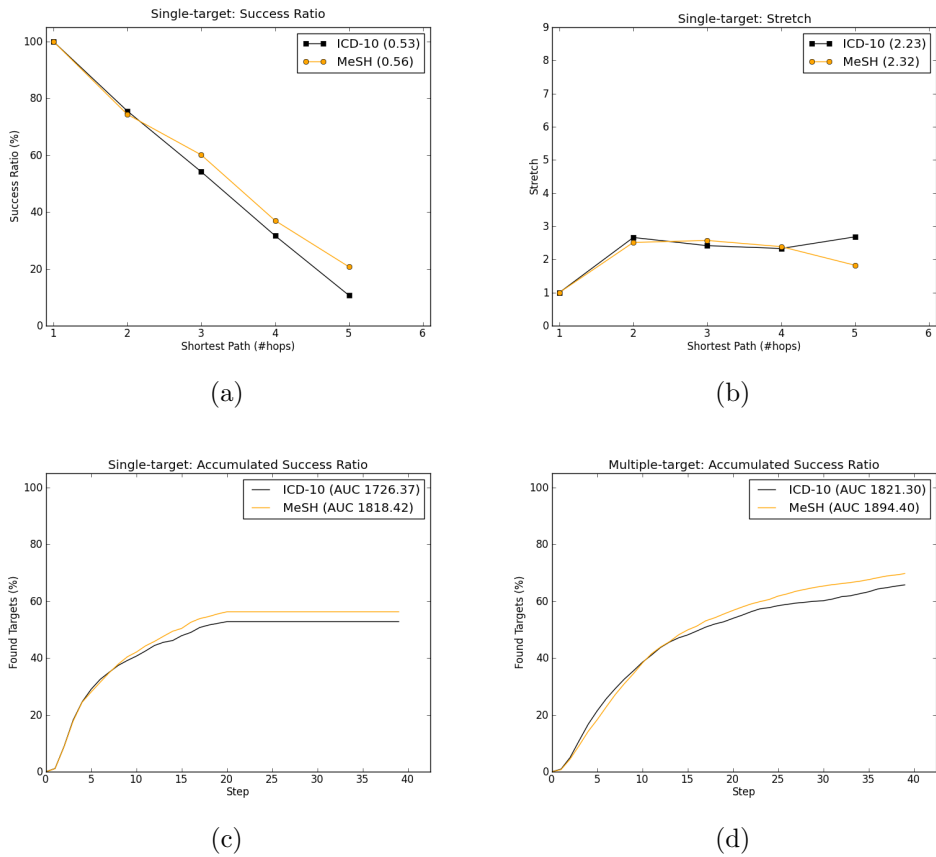


Figure 8: **Domain-specific results for ICD-10 and MeSH** The figures show *Stretch*, *Success Ratio* and *Accumulated Success Ratio* for single-target and multiple-target, respectively. The numbers in parentheses display the overall values for the success ratio and stretch, and the *area under the curve* for the accumulated success ratio.

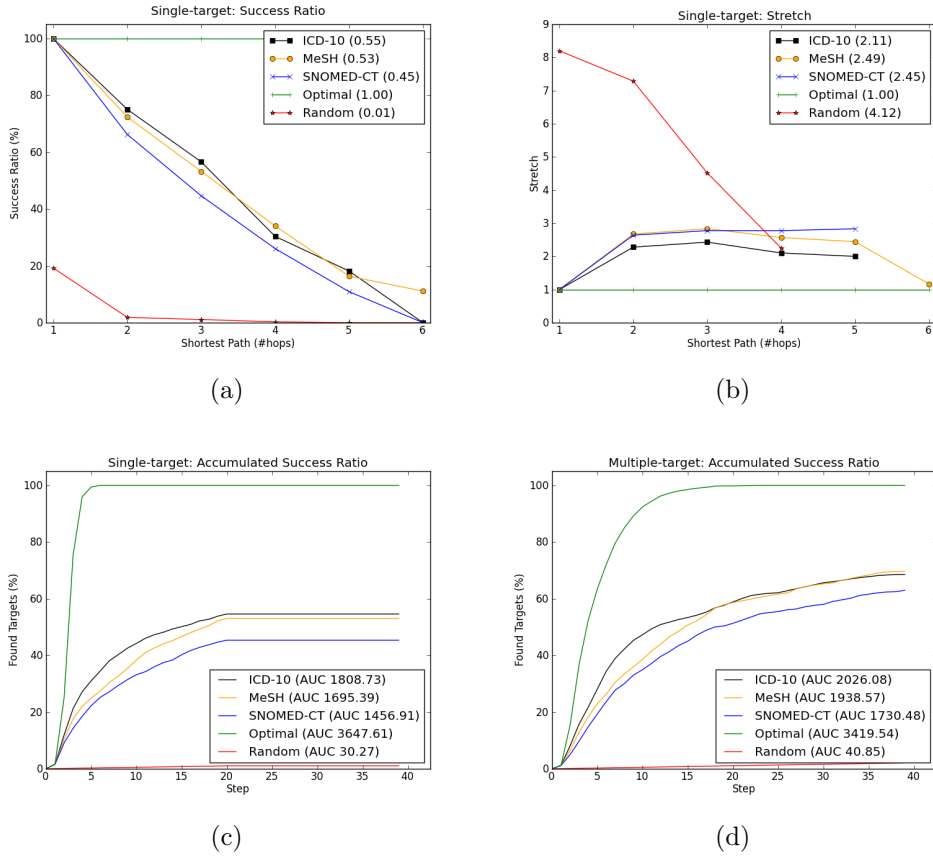


Figure 9: **Cross-domain results for ICD-10, MeSH and SNOMED-CT** The figures show *Stretch*, *Success Ratio* and *Accumulated Success Ratio* for single-target and multiple-target, respectively. The numbers in parentheses display the overall values for the success ratio and stretch, and the *area under the curve* for the accumulated success ratio. An optimal solution (green) and an averaged random walk (red) are included for comparison.

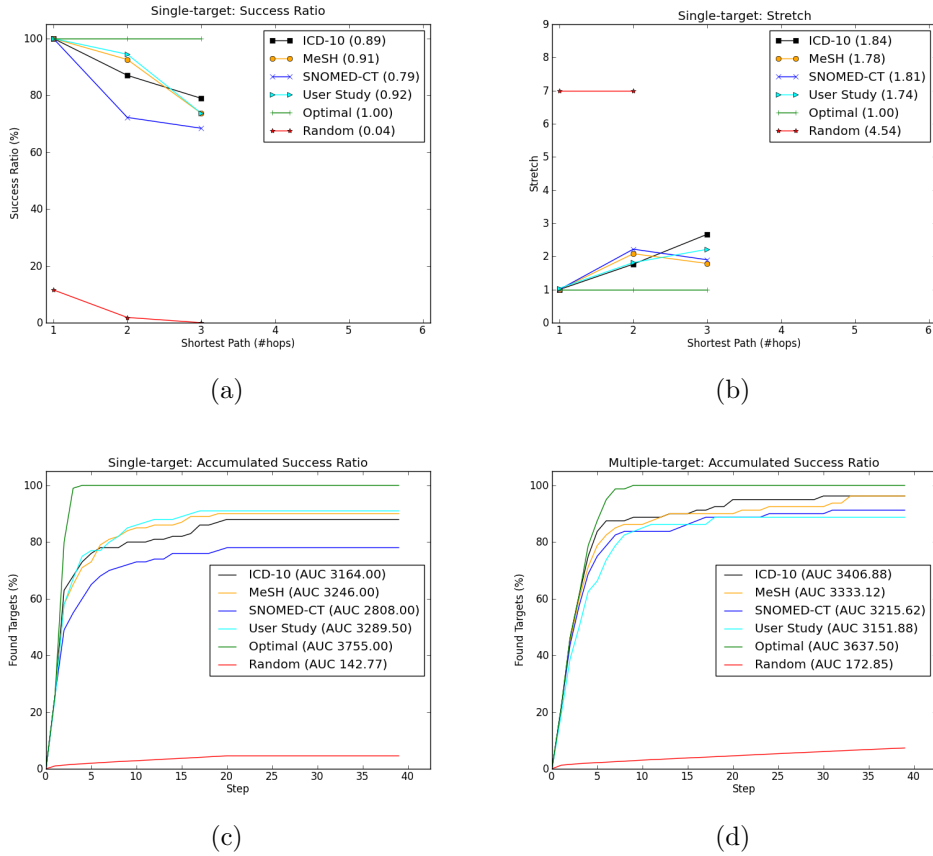


Figure 10: **Cross-domain results for the User study** The figures show *Stretch*, *Success Ratio* and *Accumulated Success Ratio* for single-target and multiple-target, respectively. The numbers in parentheses display the overall values for the success ratio and stretch, and the *area under the curve* for the accumulated success ratio. An optimal solution (green) and an averaged random walk (red) are included for comparison.

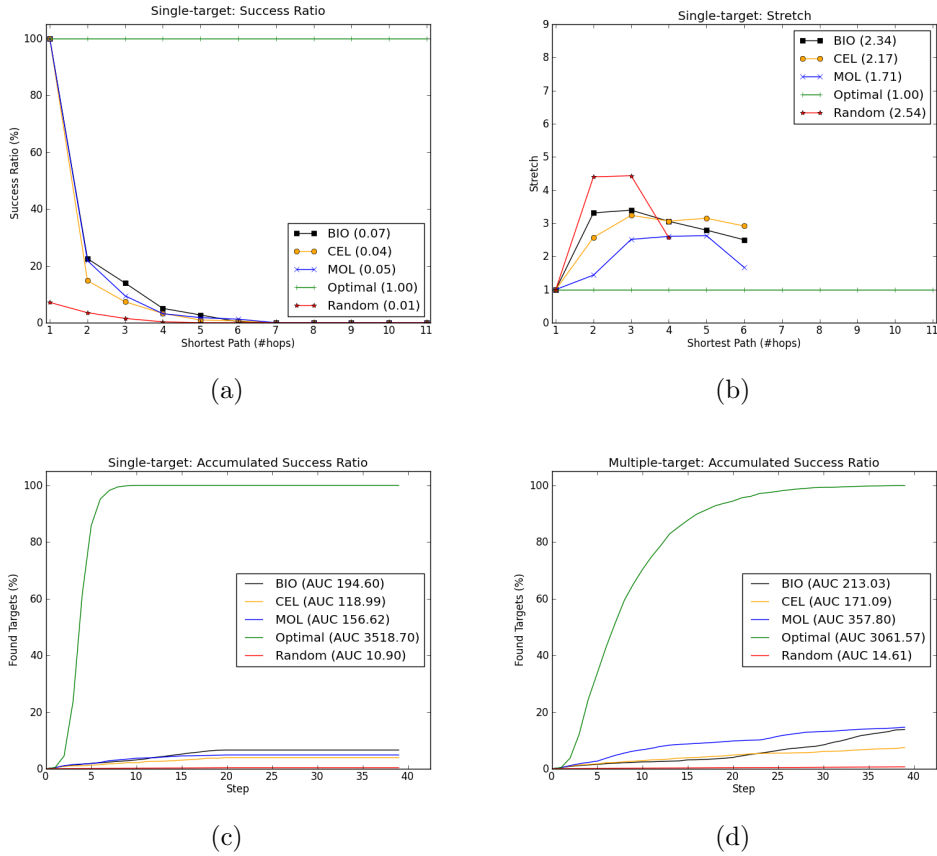


Figure 11: **Cross-domain results for the GeneOntology** The figures show *Stretch*, *Success Ratio* and *Accumulated Success Ratio* for single-target and multiple-target, respectively. The numbers in parentheses display the overall values for the success ratio and stretch, and the *area under the curve* for the accumulated success ratio. An optimal solution (green) and an averaged random walk (red) are included for comparison.

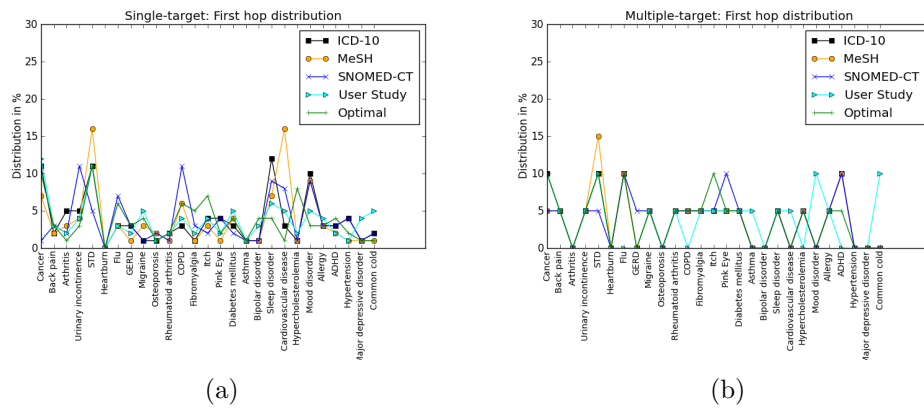


Figure 12: **Userstudy first hops** The figures show the distribution for the first clicks on the portal links for every simulation. Although technically not continuous, the data points are connected by lines to improve visibility. The data shows that the first clicks were distributed fairly evenly: No single portal link accounted for more than 17% of clicks.

5 Discussion

In comparison, the GeneOntology data set displayed a significantly lower performance than the other ontologies studied. The Wikipedia articles annotated by the GeneOntology were characterized by different properties than the other data sets: It contained a large number of stubs (very short articles) and orphans (articles not linked to by any other article). In addition, the Wikipedia article network was sparser than the other data sets, i.e., it contained fewer links (the density was 0.0013, compared to 0.0041 - 0.0057 for the other data sets). In contrast, each Wikipedia article referenced a greater number (up to fifty, in comparison to usually one or two for the other ontologies) of ontology concepts. It was hence significantly harder to discover a correct link by making an educated guess. This made navigating the graph more difficult.

In comparison to the ontologies' performance, participants in the user study performed better for single-target search and worse for multiple-target search. This is also influenced by the fact that users aborted 30% of their multiple-target navigation tasks before having found all of the targets, while the simulations ran for whole number of possible steps (40). In future work, it would be interesting to see a more extensive study without the possibility of aborting searches.

The user study was limited in that it only included "easy" target nodes, that were familiar to test subjects without a medical background. Since the simulation behavior for these targets was very close to the test subjects, we can hypothesize that behavior for the whole set of targets is likewise similar. Our evaluation method could hence be a means of replacing navigation evaluation by domain experts with automated simulation data. Again, for future work it would be interesting to compare the performances of different user groups (such as novices and medical doctors) on our data sets.

Another interesting aspect was that backtracking was the prevalent strategy for escaping dead ends and unfamiliar areas. Due to constraints in our simulation software we expected this to happen for the automatic simulations. In a first pilot study, users had specifically demanded a home button in order to directly jump back to the portal at all times. Surprisingly, users in the final study did not make use of this button very often but preferred to backtrack using the browser's back button.

6 Conclusions

With this setup we have presented an automated, task-based evaluation method for ontology structure. Our results answer several questions regarding decentralized search. First of all, we have found ontologies to be suitable to serve as background knowledge for decentralized search. With appropriate ontologies and Wikipedia link networks, the simulations produced results well above pure random walks and were able to guide navigation towards the target. Our user study, albeit limited in its extent, shows that the simulations and human behavior yielded very similar results.

A second, important finding is that our results can be used to model automatic system evaluation. We've chosen Wikipedia articles as our data basis, but our methods could, on principle, be applied to any set of documents. By using different ontologies as background knowledge for decentralized search, our method can effectively model navigation by different sets of users. By creating ontologies with different nuances of information about the document set, different types of users (such as novices and experts) could be modeled. This method could help evaluate human-computer interaction for different types of interfaces or linkage structures.

7 References

- [1] International classification of diseases, revision 10. <http://www.who.int/classifications/icd/en>, 2012.
- [2] Medical subject headings. <http://www.nlm.nih.gov/mesh/>, 2012.
- [3] M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
- [4] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. *CoRR*, abs/1111.4570, 2011.
- [5] M. Boguñá, F. Papadopoulos, and D. Krioukov. Sustaining the Internet with Hyperbolic Mapping. *Nature Communications*, 1(62), Oct 2010.
- [6] J. Brank, M. Grobelnik, and D. Mladenić. In *Proc. of 8th Int. multi-conf. Information Society*.
- [7] D. F. Gleich, P. G. Constantine, A. D. Flaxman, and A. Gunawardana. Tracking the random surfer: empirically measured teleportation parameters in pagerank. In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 381–390, New York, NY, USA, 2010. ACM.
- [8] D. Helic and M. Strohmaier. Building directories for social tagging systems. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, UK*, 2011.
- [9] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web, WWW '11*, pages 417–426, New York, NY, USA, 2011. ACM.
- [10] J. Pöschko, M. Strohmaier, D. Lamprecht, T. Tudorache, N. F. Noy, M. A. Musen. The pragmatic history behind our semantic future: Studying the evolution of large-scale ontology engineering projects and the case of ICD-11. In *Submitted for publication to the Journal of Biomedical Informatics*. Elsevier.
- [11] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing, STOC '00*, pages 163–170, New York, NY, USA, 2000.

ACM.

- [12] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá. Hyperbolic Geometry of Complex Networks. *Physical Review E*, 82(036106), Oct 2010.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [14] C. Price and K. Spackman. Snomed clinical terms. *BJHC&IM-British Journal of Healthcare Computing & Information Management*, 17(3):27–31, 2000.
- [15] M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern. Evaluation of folksonomy induction algorithms. *ACM Trans. Intell. Syst. Technol.*, 3(4):74:1–74:22, Sept. 2012.
- [16] C. Trattner, P. Singer, D. Helic, and M. Strohmaier. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *I-KNOW*, page 14, 2012.
- [17] J. Travers, S. Milgram, J. Travers, and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.
- [18] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998.
- [19] R. West and J. Leskovec. Human wayfinding in information networks. In *In WWW-12*, 2012.
- [20] R. West, J. Pineau, and D. Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.
- [21] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web-Server-Issue):541–545, 2011.
- [22] L. Yao, A. Divoli, I. Mayzus, J. A. Evans, and A. Rzhetsky. Bench-

marking ontologies: Bigger or better? *PLoS Computational Biology*,
7(1), 2011.