HHS Public Access

Author manuscript

Semant Web. Author manuscript; available in PMC 2015 November 13.

Published in final edited form as:

Semant Web. 2015 ; 6(4): 403–422. doi:10.3233/SW-140143.

Using ontologies to model human navigation behavior in information networks: A study based on Wikipedia

Daniel Lamprecht[a], **Markus Strohmaier**[a,b,*], **Denis Helic**[a], **Csongor Nyulas**[b], **Tania Tudorache**[b], **Natalya F. Noy**[b], and **Mark A. Musen**[b]

[a]Knowledge Management Institute, Graz University of Technology, Austria

[b]Stanford Center for Biomedical Informatics Research, Stanford University, USA

Abstract

The need to examine the behavior of different user groups is a fundamental requirement when building information systems. In this paper, we present Ontology-based Decentralized Search (OBDS), a novel method to model the navigation behavior of users equipped with different types of background knowledge. Ontology-based Decentralized Search combines decentralized search, an established method for navigation in social networks, and ontologies to model navigation behavior in information networks. The method uses ontologies as an explicit representation of background knowledge to inform the navigation process and guide it towards navigation targets. By using different ontologies, users equipped with different types of background knowledge can be represented. We demonstrate our method using four biomedical ontologies and their associated Wikipedia articles. We compare our simulation results with base line approaches and with results obtained from a user study. We find that our method produces click paths that have properties similar to those originating from human navigators. The results suggest that our method can be used to model human navigation behavior in systems that are based on information networks, such as Wikipedia. This paper makes the following contributions: (i) To the best of our knowledge, this is the first work to demonstrate the utility of ontologies in modeling human navigation and (ii) it yields new insights and understanding about the mechanisms of human navigation in information networks.

Keywords

Navigation; Decentralized Search; Ontology

1. Introduction

One of the challenges in building information systems is the need to develop interfaces suited to a range of different types of users. Different types of users, such as novices, experts, generalists or specialists will, in general, display considerably different knowledge about a given domain. This specific knowledge in turn influences their interactions with an information system. Gaining insight into human navigation behavior supports the

*Corresponding author – markus.strohmaier@TUGraz.at.

construction of easy-to-use software and information systems that are ready to accommodate a broad range of user types.

In this paper, we investigate ways of modeling navigational behavior of human users in information networks. Humans navigating an information network (such as Wikipedia) generally do not know the network topology in its entirety. They are therefore not always familiar with the global network structure but navigate based on assumptions and local information only. Experiments by Stanley Milgram and others [30] [20] have shown that humans are very effective at finding short paths based on local information in offline as well as in online *social networks*.

In this paper we present a novel method for simulating human navigational click behavior in information networks *using ontologies as background knowledge* and examine its suitability to model actual human navigation behavior. The method, which we call Ontology-based Decentralized Search (OBDS), builds on decentralized search [15], a well-established navigation method in social networks which is based on local information only. Decentralized search has been successfully applied to navigation in information networks in previous research, where it has been used to model the behavior of users and to produce simulated click data [11]. OBDS uses decentralized search with ontologies as background knowledge to model the search process and to point an algorithmic searcher towards the direction of the target.

This method is new in that it uses an explicit representation of the background knowledge in the form of an ontology. Research in psychology suggests that humans store concepts in their minds hierarchically [7]. In our method, we model different groups of users by using different ontologies as background knowledge.

Research Questions

In this work, we will address the following three research questions:

RQ1 Can ontologies contribute useful information to modeling navigation in information networks? And how does OBDS perform in comparison to randomly generated ontologies and random walks?

RQ2 Does Ontology-based Decentralized Search (OBDS) produce valid results, i.e., are the simulated navigation paths similar to those produced by human navigation?

RQ3 When using OBDS, what ontology is best suited to produce human-like navigation results?

To demonstrate our method, we use the information network formed by a set of biomedical Wikipedia articles and the connections (hyperlinks) between them. 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.

Contributions

Our main contribution is the demonstration of the general suitability of existing real-world ontologies to inform models of human navigation, specifically decentralized search on information networks such as Wikipedia. To the best of our knowledge, our work presents a novel method and a novel application of ontologies. By comparing the navigational paths generated by our simulations with several baseline approaches and with data obtained from a user study, we show that our method yields results similar to those produced by actual human users. The results suggest that OBDS can be used to simulate human navigational behavior in information networks, which can be useful for addressing issues arising in the development of systems that are based on networked information. These findings are relevant for researchers interested in new applications for ontologies and for researchers interested in modeling navigation in information networks using ontologies as background knowledge.

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 the results in Section 4. We end with a discussion.

2. Related work

In the context of this paper, the related areas can be divided in three: navigation in social networks, navigation in information networks and ontologies.

2.1. Navigation in social networks

This paper particularly addresses navigation in social networks via decentralized search algorithms. Fundamentally, decentralized search describes a way of solving a pathfinding problem in a social network. Starting from an arbitrary start node (i.e., a person) within the network, the objective of decentralized search is to find a way to a given target node. The algorithm, however, does not possess global knowledge of the network and can therefore only take decisions based on local knowledge. The term *decentralized* stems from the fact that the search proceeds by forwarding the search problem from one node to another, which, in a social network, involves a different person taking the decisions at every node.

The idea of decentralized search, as used in our navigation simulations, was made popular by Stanley Milgram's widely discussed small-world experiment [30] [20] 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 resulting median chain length of six intermediates for successful chains of letters coined the term "six degrees of separation". By taking only the limited knowledge of each participants into account at each step, the search effectively constituted a form of decentralized search. The result illustrated the so-called *small world phenomenon*, as it seemed possible to connect two arbitrary persons across the United States through a very small number of hops.

In 1998, Watts and Strogatz [32] characterized networks exhibiting small-world characteristics as having a high clustering coefficient and a low characteristic path length

and demonstrated the actual existence of this type of small-world networks in a film actor collaboration network, the power grid of the western United States and the neural network of C. elegans.

In 2000, Jon Kleinberg proved that for the type of small-world networks proposed by Watts and Strogatz [13], no effective decentralized search algorithm could exist that always found a path connecting two nodes in subpolynomial time. However, Kleinberg presented a more generalized version of the model for which he then proved that a decentralized algorithm capable of finding short paths existed.

Kleinberg later extended his model of decentralized search to include hierarchies [14], where the term *hierarchy* denotes a tree of that includes all network nodes (and may contain more nodes). He showed that when the network nodes were embedded as the leaf nodes of a hierarchy and links in a network were formed proportional to distances in this hierarchy, the resulting network was also efficiently searchable. To form an effectively searchable graph, nodes were connected with a probability proportional to their distance in the tree, i.e., the height of their closest common ancestor. Provided the hierarchy information as *background knowledge*, the search could then proceed to the target effectively. In this paper, we use ontologies as this type of background knowledge.

Miao et al. [19] have studied decentralized search in collaboration networks. Collaboration networks differ from information or social networks in that the information flow in them is driven by tasks. This means that the edges in the network are formed by collaboration on tasks. In their study, the tasks were software bugs. Developers who were assigned a bug they could not eliminate themselves forwarded it to another developer who they believed could handle it. By establishing several forwards in a row, this of work flow consisted a type of decentralized search, as all decisions about the next hop were taken independently by multiple participants.

Adamic and Adar [4] studied decentralized search in the e-mail network of HP labs and found that decentralized search according to hierarchies based on connectedness and office cubicle distance worked best.

Decentralized search is also used in peer-to-peer file sharing protocols such as Gnutella or KaZaA. With a low characteristic path length and a high cluster co-efficient, the Gnutella network displayed small-world characteristics in 2003 [17].

2.2. Navigation in information networks

In this paper, decentralized search, a navigation model originally developed for social networks, is applied to information networks.

One of the most prominent related model to search in information networks is information foraging [25]. Information foraging is based on foraging theory in biology. In order to survive, animals have adopted methods which maximize the energy gained from food sources. In the theory of information foraging, search in information networks is not guided by background knowledge but by information scent, with each article and link emanating a

distinct scent, which is dependent on the target of the search. For instance, when searching for information on penguins, a link leading to an article about Antarctica would provide more scent than a link leading to an article about the Sahara desert.

In this paper, information networks are studied on the example of Wikipedia. However, genuine navigation paths from Wikipedia are difficult to obtain, as the goals of users are often hidden and not explicitly visible and logs of click trails are hard to obtain. With 60 – 70%, the fraction of teleports is furthermore significantly higher on Wikipedia than on general web sites [8]. This might be due to the fact that users visit Wikipedia to satisfy specific information demands rather than to browse articles. However, there exist valid reasons to navigate Wikipedia, which will be detailed in the description of the navigation scenarios in Section 3.5.

Due to the difficulty of obtaining Wikipedia navigation paths, wiki games have been a popular replacement for Wikipedia navigation paths in recent research. Wiki games, such as Wikispeedia[1], Wikipedia Maze[2] or Wiki Game[3] allow users to play games on the information network formed by the Wikipedia articles and links between them. Click trails from wiki games have enabled researchers to gain insight into navigational behavior on Wikipedia. In 2009, West et al. [34] used wiki game data to infer semantic distances between concepts by studying game click paths. In 2012, West and Leskovec [33] found that in wiki games, players tend to navigate to hubs (articles with a large number of outlinks) first, and subsequently home in on targets node.

In our own group, we have used decentralized search (with non-ontological background knowledge) in different contexts:

In 2011, Helic and Strohmaier compared the navigability of different tag hierarchy generation algorithms on data from Bibsonomy, CiteULike, Delicious, Flickr and LastFm [10]. The paper evaluated the suitability of tag hierarchies for navigation on tagging networks and proposed a novel tag hierarchy generation algorithm.

In 2012, Strohmaier, Helic et al. compared different folksonomy induction algorithms through decentralized search [28]. They showed that, based on evaluation through navigation, clustering algorithms developed for social tagging systems performed better than standard hierarchical clustering algorithms.

Helic et al. applied decentralized search to broad and narrow folksonomies on data from Mendeley [9] and found broad folksonomies better suited to supporting navigation.

Trattner et al [29] compared decentralized search and human navigation behavior in information networks and showed that the simulation of decentralized search yielded very similar results to actual human navigation data on Wikipedia. In their work, Trattner et al. investigated different types of hierarchies as background knowledge and found that decentralized search based on a hierarchy generated from network features such as in- and

[1] www.wikispeedia.net
[2] www.wikipediamaze.com
[3] www.thewikigame.com

outdegree simulated human navigation better than comparable hierarchies generated from external knowledge.

In ongoing research, Helic, Strohmaier et al. are studying the influence of stochasticity and different methods of selecting the next hop in decentralized search [11].

The previous work did not tap into existing ontologies as background knowledge, but used other approaches (such as automated methods) for this purpose. This paper goes beyond previous research by extending the simulation framework with ontologies and by applying Ontology-based Decentralized Search to the case of Wikipedia and for concrete ontologies for the biomedical domain.

2.3. Ontologies

Ontologies have been used in previous research to facilitate navigation in digital libraries. Papazoglou and Hoppenbrouwers [23] have used ontologies to retrieve related work when searching digital libraries. The research of Rajapakse et al. [27] shows efforts to navigate the digitally available literature related to dengue fever. Villela Dantes et al. [31] have studied the ontology-guided insertion of links into web pages. In their work, they classified web pages according to an ontology and subsequently inserted links to related topics into web pages to facilitate navigation.

These research papers share the effort to use ontologies to aid navigation. The objective of this paper lies in explaining and modeling user behavior by using ontologies as background knowledge. The ontologies are hence not used to guide human users but to simulate and possibly explain behavior.

This paper uses three ontologies from the biomedical domain. Biomedical ontologies play an important role in biomedical research [6] and are used for a range of purposes. In the biomedical domain, ontologies have been adapted more frequently than in other disciplines [22].

3. Materials and methods

3.1. Introductory example

To illustrate our work, let us introduce the following example, depicted in Figure 1: *Alice accompanied her father to a physician, who diagnosed him with a certain cardiovascular disease. Back at home, Alice realizes that she forgot the exact name of the condition. However, she remembers that the disease was somehow related to heart rhythm problems. Trying to recover the exact name, she goes to Wikipedia, but since she does not know the exact name of the target article she cannot use the search function to jump to the article directly. Alice instead starts from a (hypothetical) Wikipedia portal containing links to a number of common diseases. She first chooses to click the portal link leading to the article on Cardiovascular disease, as this seems to be a good starting point. Next, she navigates to the article on Vascular disease, then to Stroke, clicks the link to Cardiac dysrythmia and finally arrives at Supraventricular tachycardia, which she recognizes as the disease the doctor had diagnosed her father with.*

At each step, Alice is only aware of the links leading away from the current article. She is familiar with some of the article titles, and is able to relate them to one another through what we refer to as her *background knowledge*. She recognizes some of the links and knows what their target article could 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*.

To simulate Alice's usage of Wikipedia, we first mapped a subset of biomedical Wikipedia articles to their corresponding ontology concepts in three biomedical ontologies. Given these mappings, the simulation could then calculate distance information on the ontology. For each potential outgoing link that Al-ice could click, the simulation computed the shortest path between the article behind that link and the target article. This distance information was used as a proxy measure to estimate the distance to the target article in the Wikipedia network. The distance information gained this way was not necessarily optimal or even correct, but generally provided a good guess to guide the navigation.

In this manner, the simulator was able to make an educated guess about what link to follow, just as Al-ice could roughly place the outgoing articles into categories.

In the rest of the Materials and Methods section, we describe the ontologies used to inform the simulator (Section 3.2), the Wikipedia articles and how we obtained them (Section 3.3), our method of Ontology-based Decentralized Search (Section 3.4), our navigation scenarios (Section 3.5), the user study (section 3.6) and finally the simulator implementation (section 3.7).

3.2. Biomedical ontologies

We used the following four ontologies and terminologies (all from the biomedical domain) as background knowledge. With these ontologies, we were able to i) extract articles from Wikipedia and ii) guide the next-step selection in the simulator.

The **International Classification of Diseases, 10th 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. In our experiments, we used Wikipedia articles mapping to concepts from all 22 chapters.

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 graph we used contained 80,689 nodes. MeSH extends beyond biomedical concepts and comprises terms from other domains such as Geography, Technology or Publication Characteristics. In our experiments, 96% of

the Wikipedia articles mapped to the subgraph represented by the *Diseases* concept, and the rest to the *Psychiatry and Psychology* subgraph.

The **Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT)** [26] is a clinical health-care terminology used in electronic health record 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 our experiments, 98% percent of the Wikipedia articles mapped to the *Clinical finding* subtree.

Table 1 displays statistics about the data sets used for this paper. The row denoted *density* was calculated as

$$D=\frac{|\text{relations}|}{|\text{concepts}|(|\text{concepts}|-1)}$$

for the ontologies (which were regarded as undirected graphs) and as

$$D=\frac{|\text{links}|}{|\text{articles}|(|\text{articles}|-1)}$$

for the Wikipedia article network, which formed a directed graph. Figure 2 depicts the examined ontology graphically for the first four hops from the root node.

3.3. Wikipedia articles

We used a dump of the English Wikipedia from December 2011 to extract articles from the biomedical domain corresponding to ontology concepts. We then mapped the articles to the ontologies by parsing the articles' info boxes.

In disease articles, the Infobox disease[4] is commonly used. It offers several options to reference medical ontologies such as ICD-10 or MeSH (see Figure 3 for an example). We used template fields in the Infobox disease as well as two 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 Bio-Portal [35] to map Wikipedia articles to SNOMED CT. We mapped a total of 1,593 Wikipedia articles from both ICD-10 and MeSH to SNOMED CT with this method.

3.4. Ontology-based Decentralized Search

Decentralized search is a method of solving a pathfinding problem in a network without a central control unit. Starting from an arbitrary start node within the network, the objective of

[4] http://en.wikipedia.org/wiki/Template:Infobox_disease

decentralized search is to find a way to a given target node. The term *decentralized* stems from the fact that the search proceeds by forwarding the search problem from one node to the next, until the target is reached. In Stanley Milgram's small world experiment [30], decentralized search was established through humans forwarding letters to acquaintances in order to find a target person. Each human along the chain of letters acted independently of all others and thus made the search decentralized, i.e., acting without a central control unit involved in the decisions at every step. Further examples for decentralized search include bug forwarding in a developer network, where software bugs are assigned to a starting person, and then forwarded to other developers until it is fixed [19], or job recommendations in social networks [4].

In a social network, the decision of where to forward the problem is generally based on the expected knowledge and capability of that particular next node (person). For our simulations, we assumed that all nodes shared a common background knowledge expressed as an ontology. This assumption made our algorithm less "decentralized" in a certain sense, because all the decisions were now made by the same entity (our simulator). Just like in the original decentralized search however, at each node the simulator could only access information about that particular node's local network neighborhood. The background knowledge represented additional knowledge about the network necessary to effectively find a short path to the target. When looking for an employee in a company for example, this knowledge could represent the organizational hierarchy - with the restriction that the search can only be forwarded to acquainted employees, which would e.g., be the case with personal recommendations.

In the theory of network navigability, Jon Klein-berg showed that networks that are formed according to a background hierarchy (i.e., a tree) are efficiently navigable [14], provided the search agent has access to that background hierarchy during the search. This method, called Hierarchical Decentralized Search, has been successfully applied in previous research [11] [28]. This paper extends this application by a using ontologies as the background knowledge.

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 (see Figure 1 for an example).

To avoid loops, the simulation explored each node in the network only once. However, the simulation could backtrack to the last visited nodes (up until the starting portal, if necessary), just as Alice would use her browser's back button. This was used in case of dead ends (articles with no unvisited outgoing links) or at articles providing only links leading further away from the target (according to the ontology information). At any given point, the simulation could also jump back to the starting portal directly, modeling a home button in an information system.

This use of existing ontologies represents a substantial change in the motivation of the background knowledge: As opposed to previous work in this area, the background knowledge is now exogenous to the network. What this implies is that the hierarchy is based on knowledge independent of the network that the agent navigates on. All ontologies used in the application of Ontology-based Decentralized Search in this paper play a key role for their corresponding domain in their research fields. They are hence representative for a good part of the knowledge in these domains. This provides OBDS with a foundation to more accurately represent the intuitions of human navigation behavior.

The use of ontologies and the associated semantic information open up a range of new possibilities for the application of the background knowledge:

- *Filtering by relations and properties:* Ontologies are (in general) made up of different types of relations (such as *is-a* or *part-of*, or *regulates*), which can be used to extract different varieties of background knowledge from one and the same ontology. For example, a hierarchical version of the ontology could be extracted by following only the *is-a* relations. Furthermore, ontologies may assign *properties* to their concepts. A background knowledge can hence also be restricted to ontology concepts with a certain property. An ontology could be filtered to contain only contain concepts stemming from a single domain, such as *geography*. This could then be compared with other filtered versions of the ontology.
- *Modeling different user groups:* Ontologies can also be used to model different types of users. A good example for this is the case of ICD-10, which provides a classification of diseases. In the ontology, the depth of a disease (i.e., its distance from the root node) corresponds to its specificity. This could be used to model the knowledge of different hospital personnel. For instance, a medical specialist could be modeled by the entire depth of knowledge of one section of the ontology, and a depth-limitation in the other sections. A layperson could be modeled by having a certain depth-limitation in all areas. This could be effectively used to simulate different user groups in medical information systems, without having to carry out actual human user studies.
- *Inference:* Ontologies permit *inference* on their entities. For hierarchical relations this could mean that subconcepts could be assigned the type of their superconcepts (e.g., the perhaps unfamiliar Supraventricular tachycardia is a subconcept of Heart Disease in ICD-10, which is more commonly known). In the case of the cut-off background knowledge, more specific ontology concepts could then be substituted by their inferred superconcepts and provide more information to the navigation process than a pure random guess.

In the experiments conducted for this paper, Ontology-based Decentralized Search was used with three different ontologies (ICD-10, MeSH and SNOMED-CT) that were not filtered. This meant, that all concepts and relation types present in the ontologies (and mapping to the data sets) were used as the background knowledge. The results of the simulations were then compared on the same information network, which meant that the three ontologies effectively modeled different user groups on the same set of data. While all three ontologies

represent expert knowledge, they still server different purposes: ICD-10 is a disease classification widely used by insurance companies, physicians and hospitals. SNOMED-CT is a terminology of clinical terms, and MeSH is a terminology for the indexing of journal articles. We hence assume these ontologies to be representative of experts in their respective fields.

3.5. Navigation scenarios

We studied two different search scenarios, both of which started from a hypothetical Wikipedia portal.

Starting Portal—We started the navigation from a hypothetical Wikipedia portal featuring a selection of suitable articles: The 25 health conditions listed in the main navigation toolbar of WebMD.com (see Figure 4). We manually mapped these conditions to Wikipedia articles from our dataset and used the articles as the outgoing links from our artificial portal. In a way, the artificial portal thereby resembles the navigational structure of the WebMD front page - a popular health information web site. Medical web sites, such as WebMD are frequently [5] used to obtain information about diseases or as a first information before consulting a medical doctor.

Single-target search—Our first scenario was analogous to Alice's introductory example. In single-target search, the simulation started at the portal and proceeded to a single target article using Ontology-based Decentralized Search.

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

Multiple-target Search—For multiple-target search, the difference was in the targets, which consisted of target sets of two to ten articles (instead of only a single one). 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 [18].

We used clusters of semantically similar Wikipedia articles as our target sets and applied *k-means clustering* to arrange similar articles into clusters based on *TF-IDF* features (using scikit-learn [24]). We used those resulting clusters containing two to fifteen articles in our simulations. Examples for clusters are given in Table 2.

3.6. 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. All of them were graduate students in different fields (but not in medicine) at Stanford University at the time of the user study.

The study used the data set of ICD-10, SNOMED CT and MeSH, containing 1,593 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, which left users puzzled in a pilot study), 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 our testing environment containing the subset of Wikipedia, 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 home link were enabled at all times.

The setup for the user study consisted of a web site similar to Wikipedia, which contained the articles used the study, as well as information about the current task. Each step of the user was logged. This setup is visible in Figure 4.

Each participant completed a total of 15 navigation tasks. A navigation task consisted of finding a given target node (or a set of target nodes) in the subset of the Wikipedia network. As in the simulator, the starting point for a task was always the portal, and participants could only click on links to articles within the data set. 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).

3.7. Implementation

The experiments presented in this paper were conducted on a decentralized search simulator. This simulator was an extension of previous work by Helic, Strohmaier et al ([10], [28]) and implemented in C++ based on the Stanford Network Analysis Project framework [3]. It permitted the simulation of decentralized search on a given network and used a provided background knowledge to calculate the distances. The simulator was used to run a total of 1794 simulations of decentralized search with two navigation scenarios.

4. Results

4.1. Evaluation metrics

Based on work by Krioukov and Papadopoulos [16] we used *success ratio* and *stretch* to evaluate navigation paths.

In accordance with Strohmaier and Helic [11], we define *success ratio s* to be the fraction of target nodes found and *stretch* τ to be the average ratio of found path lengths to shortest path lengths.

Let P be the set of target nodes and W be the set of target nodes that were successfully navigated to by our simulator. Then we have that the success ratio s is

$$s=\frac{|W|}{|P|}.$$

Thus, the success ratio measures the extent to which the simulator is successful in finding a target, e.g., a success ratio of 90% states that 90% of the targets have been found. Furthermore, let $l(t)$ be the length of the shortest path from the portal to the target node t and let $h(t)$ be the length of the path to the target found by the agent. The stretch τ is then defined as

$$\tau = \frac{1}{|W|} \sum_{t \in W} \frac{h(t)}{l(t)}.$$

Stretch measures the efficiency of search. For example, a stretch of 1.2 states that the paths an agent was able to find are - on average - 20% longer than the shortest paths for these targets. As in work by Helic, Trattner et al ([10], [29]), we report success ratio and stretch split by 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 fraction of nodes found up until a certain number n of steps.

$$as(n) = \frac{|W_n|}{|P|},$$

where W_n is the set of target nodes reached by the simulation in n steps or less.

For all our evaluations, we assumed a maximum number of 20 clicks for the single-target scenario and 40 clicks for the multiple-target scenario.

4.2. Comparison with random baselines and optimal solutions

We established comparisons with random and optimal solutions by including a random walk, randomly generated ontologies and a shortest-path solution.

Random Walk—The random walk consisted of following a random link (or tracking back) at each step, not taking already visited nodes or potential targets among the neighboring nodes into account. The comparison with the random walk showed us how much more information the OBDS approach provided to the navigation in comparison to a completely random behavior.

Randomly generated Ontologies—For this comparison, we constructed a randomly generated ontology counterpart to each ontology used in our simulations. To this end, we used the number of nodes and edges as input for the configuration model approach of generating a random graph with the same number of nodes and edges [21]. As the resulting graph was not necessarily connected, we subsequently randomly connected all graph components and then removed the number of additional edges created in this process from other parts of the resulting graph (without deconnecting it).

This comparison showed us how much information the OBDS approach gained by taking the structure of the ontologies into account (but not yet the correct mappings). Furthermore, evaluating with randomly generated ontologies took the structured search behavior of decentralized search into account: Decentralized search, in our implementation, did not re-explore already visited nodes, could backtrack and always recognized links leading to a target node among the current node's neighbors. This gave this method a distinct advantage over the pure random walk.

Shortest-path solution—Finally, for the optimal solution, we computed a shortest-path solution. In the single-target scenario, this meant that we always used the shortest possible path in the graph for connecting the portal to the target node. For the multiple-target scenario, an exact solution would have required solving an instance of the traveling-salesman problem, which is computationally expensive. To circumvent this issue, we approximated the perfect solution with a nearest-neighbor approach that always took the shortest possible path to the nearest neighbor. This allowed the us to compare to the (approximately) optimal solution. It is important to note that this was only possible with global knowledge of the graph topology, which users do not posses in a decentralized search scenario.

4.3. Evaluation

To compare the performance of OBDS with different ontologies as background knowledge, we evaluated multiple ontologies on the same set of Wikipedia articles. This allowed us to inspect multiple ontologies side by side, facilitating comparison.

The results (Figure 5) show that the success ratios were well above both the random walk and the randomly generated ontologies. When comparing OBDS with different ontologies, the results show that OBDS with ICD-10 performed best, followed by MeSH and SNOMED CT for the success ratios. For the stretch, SNOMED CT fared slightly better than MeSH (with an average stretch of 2.45 resp. 2.49).

4.4. User study

For the user study, we compared the results of human navigators with OBDS. The targets were 100 manually selected targets and 20 manually selected clusters (which were the same for both the users and the simulator). This limitation of targets also meant that targets were a maximum distance of three hops away from the portal. The evaluations hence do not include any data points for longer shortest paths.

Figure 5 shows that the success ratios for the user study were fairly close to the simulator. For the single-target scenario, the overall success ratio was 92% for the user study and ranged from 79 - 91% for the ontologies. For the multiple-target scenario, the accumulated success ratio shows that the user study fell within or just below the range of the three ontologies. It is worth noting that after 20 steps, the users in our study did not find any more targets. This coincides with the point from where on users where given the possibility to abort a search task if they could not find the target. For the single-target stretch again, with

an overall stretch of 1.74 the user study performed slightly better than the simulator, which displayed stretches between 1.78 and 1.84.

To obtain qualitative insight into the navigation process, we compared the produced path lengths of the user study and the simulator. To this end, we examined the distribution of path lengths produced by both the user study and the simulations. This distribution can be seen in Figure 6. We then computed the Kullback-Leibler divergence from the user study distribution to the other distributions. The Kullback-Leibler divergence measures the number of additional bits needed to encode the path length distribution, if the other distribution is used in place of the original (user study) path length distribution. The resulting values can be seen in Table 3. For the single-target search scenario, it is clearly visible that only OBDS with produced path length distributions close to the user study: All three ontology path length distributions had a very small KL divergence (0.08 – 0.18 bits) to the user study. This means that it is justifiable to replace human navigation data with data produced by OBDS and a fitting ontology (as far as produced path lengths are considered). The same cannot be said about randomly generated ontologies (nor the random walk or the optimal solution), which cannot be easily taken in lieu of the ontologies and yield similar results.

For the multiple-target search scenario, this assertion cannot be made this clearly. However, the path length distribution for the multiple-target scenario was rather sparse, as there were merely twenty search scenarios, all of which were very likely to produce a path of a different total length. This meant that a single path accounted for five percent of the path lengths, which is also reflected in Figure 6b).

In addition, we analyzed several further aspects of the user study in comparison with the ontologies, displayed in Table 4. First, we looked into the first visited nodes and the found targets. To compare these, we arranged the nodes into vectors and computed cosine similarities.

For the found targets, all three ontologies displayed high cosine similarity values. This reflects the results from Figure 5, and is caused by 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.

For the first hops (i.e., the very first clicks in the search), the clicks were distributed rather evenly. A truly random distribution would see each link clicked 3.7% of times. Our results showed distributions ranging from 1 to 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.

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. 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.

5. Discussion

In this work, we studied simulated user navigation behavior via decentralized search. We introduced Ontology-based Decentralized Search (OBDS), a novel navigation simulation method based on decentralized search which uses ontologies as background knowledge. We showed that our method can be successfully applied to navigation in information networks, and demonstrated it can be applied on of Wikipedia supported by biomedical ontologies.

In the following, we want to focus our discussion on the research questions raised in this work:

RQ1 Can ontologies contribute useful information to modeling navigation in information networks? And how does OBDS perform in comparison to randomly generated ontologies and random walks?

We found that ontologies can indeed inform navigation in information networks. OBDS with medical ontologies as background knowledge was able to outperform the random baseline approaches significantly.

RQ2 Does Ontology-based Decentralized Search (OBDS) produce valid results, i.e., are the simulated navigation paths similar to those produced by human navigation?

We addressed this question by comparing certain properties of the simulated navigation paths with properties produced by humans in a study. We found that the click paths produced by OBDS matched certain properties of human paths better than pure random walks and randomly generated ontologies.

RQ3 When using OBDS, what ontology is bested suited to produce human-like navigation results?

From our results, it seems that ICD-10 and MeSH seem to perform best. However, the overall differences between the ontologies were not very strong, and it is subject of ongoing research to further identify differences in the performance of OBDS with different ontologies.

5.1. Further comments

We've limited our work to ontologies and Wikipedia articles from the biomedical domain in this paper. In this domain, ontologies have been adapted more frequently than in other disciplines [22], play an important role in biomedical research [6] and are used for a range of purposes. Another important aspect was the ready availability of infobox templates on Wikipedia articles, which facilitated the mappings to the ontologies. However, the principles of our method apply for other domains as well.

Influence of ICD-10—The International Classification of Diseases (ICD-10) has found widespread use and probably influences and inspires Wikipedia editors. On Wikipedia, disease articles are almost always indexed by ICD-10 as the first entry in the article infoboxes. Furthermore, the category system for the disease articles of the English Wikipedia is organized according to ICD-10. These two facts and the wide use of ICD-10 have quite possibly also influenced the link creation behavior on the encyclopedia as well as the general knowledge of the test subjects. This might be an explanation of why ICD-10 seems to be best suited to model human navigation behavior in our case study.

User Study—In comparison to the simulator's 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.

Building User Models—By using different ontologies as background knowledge, our results could help researchers and engineers build and evaluate user interfaces with different user types. The ontologies compared in the results were rather similar and mostly shared the same domain. In future work, it will be interesting to compare ontologies that do not cover the entire domain, modeling specialist users, or combining ontologies to form a more complete coverage of the domain. Another idea might be to prune the ontologies at a certain depth, modeling broad generalist knowledge that does not extend beyond a certain depth.

Action Selection—The simulations in the form we presented followed a deterministic greedy action selection model, in that it always selected the most profitable link according to the background knowledge. Related research has shown that users might be better modeled using epsilon-greedy action selection mechanisms with dynamically changing epsilons [11]. In follow-up research, our work could be extended with stochastic action selection mechanism such as epsilon-greedy. This would also lead to another potentially crucial aspect of the present simulations, namely the need to evaluate games multiple times with potentially varying results. One could expect that these adaptations would help to finetune and validate any future attempts at modeling human navigation. However, we leave this task to future research.

5.2. Future work

The user study we presented was limited in that it was restricted to a subset of target nodes because of the requirement that the target should be familiar to test subjects without a medical background. Since the simulation behavior for these targets was very close to the

test subjects, it can be hypothesized that the user behavior for the whole set of targets is likewise similar. It is up to future research to show more details of the comparison of human users and decentralized search.

Another aspect was the limitation of the user study to a subset of the target articles. The user study tried to approximate non-medical students with expert biomedical ontologies. While this worked to a certain extent, it will be interesting to see further user studies with medical experts and compare their results on the entire data.

The chosen portal, based on WebMD.com, undoubtedly influenced the navigation results. It is up to future work to compare different portals and shed a light on possible differences.

The idea to navigate to one single predefined target might seem somewhat artificial in the case of user behavior concerning explorative tasks. However, one idea to improve on this might be calculate the TF-IDF features of the target node beforehand and subsequently navigate until a page (or a number of pages) similar enough to the TF-IDF features has been found (which does not need to be the predefined target page). This could model the case of users exploring areas of the network.

Other potential research questions might include the limitation of visible links to links in the upper part of Wikipedia articles and comparing the results on non-English editions of the encyclopedia. Past research [12] has already compared different methods of extracting background knowledge from the actual network used for navigation. These background knowledges were based on network features such as centrality or degree. It would be interesting to directly compare these extracted background knowledges with ontologies and analyze the differences.

6. Conclusions

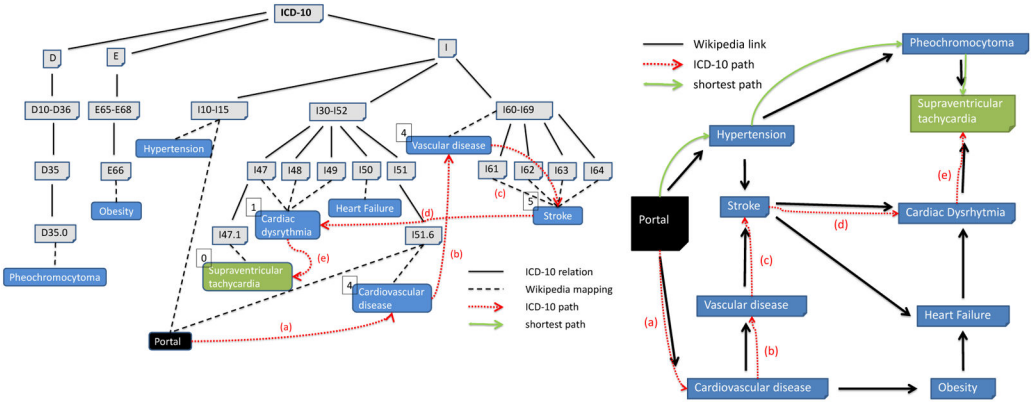
In this work, we have presented a novel, ontology-based method (Ontology-Based Decentralized Search) for simulating human navigation in information networks such as Wikipedia. Our results provide technical answers to several questions regarding the use of ontologies in decentralized search: We have not only presented a method to integrate ontological background knowledge into decentralized search, but also found that ontologies can serve as *efficient* background knowledge. With appropriate ontologies and Wikipedia link networks, our simulations using OBDS (i) found targets more efficiently than two baseline approaches (random walks or randomly generated ontologies) and (ii) produced navigational paths that were more similar to actual human navigational paths than to baseline approaches. While our human subject study was limited in terms of - for example - size, the results reported in this paper are encouraging in several ways. First, our method opens up ways to explore the effects of assuming different kinds of background knowledge of users in a navigation task. For example, swapping different kinds of ontologies in future work would allow us to explore their impact on the efficiency of decentralized search in information networks. Second, our results can be seen as additional corroboration that ontologies indeed capture useful knowledge about a domain. In some of our experiments, OBDS with medical ontologies as background knowledge was able to outperform baseline approaches significantly.

Summarizing, our findings are relevant for researchers interested in new applications for ontologies or interested in modeling navigation in information networks using ontologies as background knowledge.

References

1. [last accessed 22-April-2013] International classification of diseases, revision 10. 2012. http://www.who.int/classifications/icd/en
2. [last accessed 22-April-2013] Medical subject headings. 2012. http://www.nlm.nih.gov/mesh/
3. [last accessed 22-April-2013] Stanford network analysis project. 2012. http://snap.stanford.edu
4. Adamic, LA.; Adar, E. How to search a social network. Nov. 2004
5. Baker L, Wagner TH, Singer S, Bundorf M. Use of the internet and e-mail for health care information: Results from a national survey. JAMA. 2003; 289(18):2400–2406. [PubMed: 12746364]
6. Bodenreider O, et al. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008; 67:79.
7. Fang, J.; Evermann, J. Evaluating ontologies: Towards a cognitive measure of quality. Proceedings of the 16th IEEE International Enterprise Distributed Object Computing Conference Workshops, EDOC'07; IEEE Computer Society; 2007. p. 109-116.
8. Gleich, DF.; Constantine, PG.; Flaxman, AD.; Gunawardana, A. Tracking the random surfer: empirically measured teleportation parameters in pagerank. Proceedings of the 19th International Conference on World-wide Web, WWW '10; New York, NY, USA: ACM; 2010. p. 381-390.
9. Helic, D.; Körner, C.; Granitzer, M.; Strohmaier, M.; Trattner, C. Navigational efficiency of broad vs. narrow folksonomies. Proceedings of the 23nd ACM Conference on Hypertext and Hypermedia, HT'12; 2012;
10. Helic, D.; Strohmaier, M. Building directories for social tagging systems. Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM'11; 2011;
11. Helic, D.; Strohmaier, M.; Granitzer, M.; Scherer, R. Models of human navigation in information networks based on decentralized search. Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT'13; 2013;
12. Helic, D.; Strohmaier, M.; Trattner, C.; Muhr, M.; Lerman, K. Pragmatic evaluation of folksonomies. Proceedings of the 20th international Conference on World-Wide Web, WWW'11; ACM; 2011. p. 417-426.
13. Kleinberg, JM. The small-world phenomenon: an algorithm perspective. Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, STOC '00; New York, NY, USA: ACM; 2000. p. 163-170.
14. Kleinberg, JM. Small-world phenomena and the dynamics of information. Proceedings of the 20th Conference on Neural Information Processing Systems, NIPS'12; 2001. p. 431-438.
15. Kleinberg, JM. Complex networks and decentralized search algorithms. Proceedings of the International Congress of Mathematicians, ICM'06; 2006. p. 1019-1044.
16. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M. Hyperbolic Geometry of Complex Networks. Physical Review E. Oct.2010 82(036106)
17. Li, Z.; Zhao, X.; Huang, D.; Huang, J. An improved network broadcasting method based on gnutella network. In: Li, M.; Sun, X-H.; Deng, Q.; Ni, J., editors. Proceedings of the 2nd Grid and Cooperative Computing, volume 3033 of GCC'03. Springer; 2003. p. 404-407.
18. Marchionini G. Exploratory search: from finding to understanding. Commun ACM. Apr; 2006 49(4):41–46.
19. Miao, G.; Tao, S.; Cheng, W.; Moulic, R.; Moser, LE.; Lo, D.; Yan, X. Understanding task-driven information flow in collaborative networks. Proceedings of the 21st International Conference on World-Wide Web, WWW '12; New York, NY, USA: ACM; 2012. p. 849-858.
20. Milgram S. The small world problem. Psychology Today. 1967; 1(1):61–67.
21. Newman MEJ. The structure and function of complex networks. SIAM Review. 2003; 45(2):167–256.

22. Noy, NF.; Tudorache, T. Collaborative ontology development on the (semantic) web. Proceedings of the AAAI Spring Symposium on Semantic Web and Knowledge Engineering; 2008;
23. Papazoglou M, Hoppenbrouwers J. Knowledge navigation in networked digital libraries. Knowledge Acquisition, Modeling and Management. 1999:13–32.
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.
25. Pirolli, P. Information Foraging Theory: Adaptive Interaction with Information. Oxford University Press; 2007.
26. Price C, Spackman K. Snomed clinical terms. BJHC&IM-British Journal of Healthcare Computing & Information Management. 2000; 17(3):27–31.
27. Rajapakse M, Kanagasabai R, Ang WT, Veeramani A, Schreiber MJ, Baker CJ. Ontology-centric integration and navigation of the dengue literature. Journal of biomedical informatics. 2008; 41(5): 806–815. [PubMed: 18524685]
28. Strohmaier M, Helic D, Benz D, Körner C, Kern R. Evaluation of folksonomy induction algorithms. ACM Transactions on Intelligent Systems and Technology. Sep; 2012 3(4):74:1–74:22.
29. Trattner, C.; Singer, P.; Helic, D.; Strohmaier, M. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, IKNOW'12; 2012. p. 14
30. Travers J, Milgram S. An experimental study of the small world problem. Sociometry. 1969; 32:425–443.
31. Villela Dantas JR, Muniz Farias PP. Conceptual navigation in knowledge management environments using navcon. Information processing & management. 2010; 46(4):413–425.
32. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature. Jun; 1998 393(6684):440–442. [PubMed: 9623998]
33. West, R.; Leskovec, J. Human wayfinding in information networks. Proceedings of the 21st international Conference on World-Wide Web, WWW'12; ACM; 2012. p. 619-628.
34. West, R.; Pineau, J.; Precup, D. Wikispeedia: An online game for inferring semantic distances between concepts. Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI'09; 2009;
35. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. 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. 2011; 39(Web-Server-Issue):541–545.

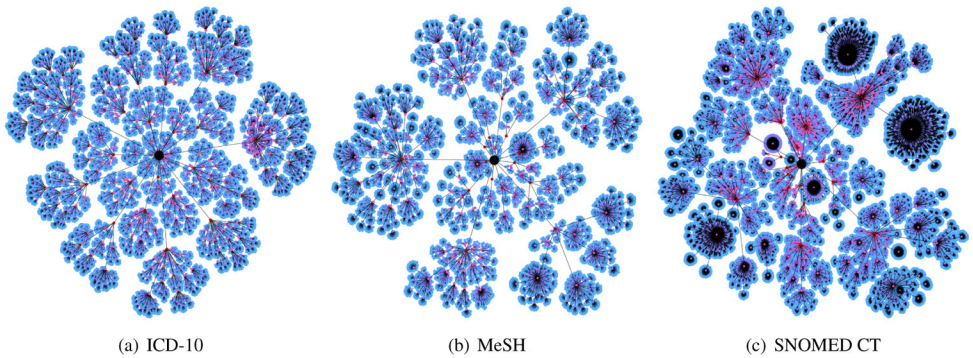


(a) Background knowledge (ICD-10) (b) Wikipedia graph

Fig. 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.

In the figure, we assume that Alice's background knowledge is represented by the ICD-10 ontology. Figure a) shows a part of ICD-10 and the corresponding Wikipedia articles. Figure b) shows a subgraph of the Wikipedia link network. Alice's path in the graph (red, dashed) is guided by ICD-10, which differs from the shortest path (green, solid). The numbers along the ICD-10 path show the distance to the target, according to ICD-10.



(a) ICD-10 (b) MeSH (c) SNOMED CT

Fig. 2. Structure of the four top levels of the ontologies used in our research

The figure shows the structure for ICD-10, MeSH and SNOMED CT. The root node is displayed in black and bold in the middle of each plot. The figures show all ontology concepts up until a distance of four from the root node. 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 more concepts than the former.

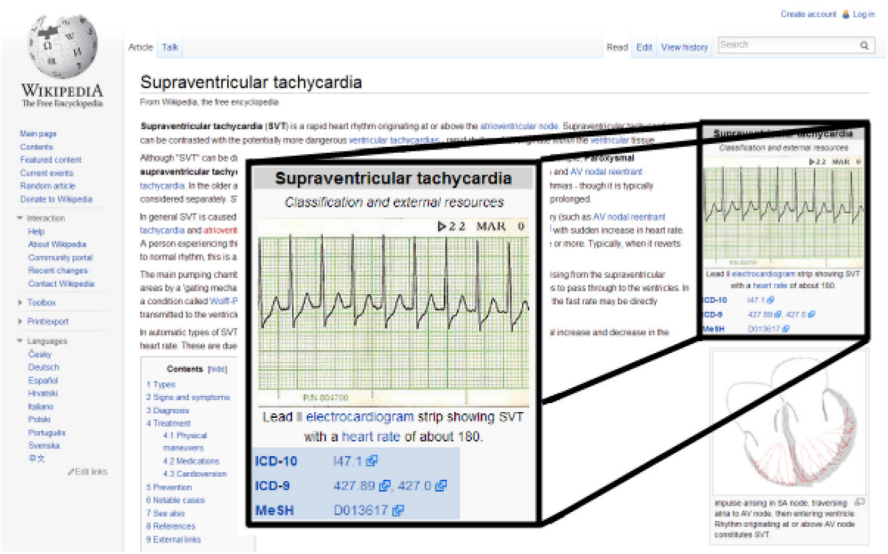
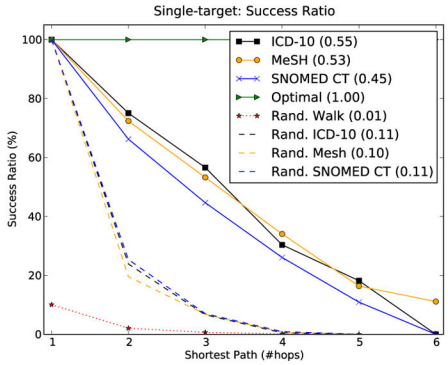


Fig. 3. Example for an infobox template used in disease articles on Wikipedia

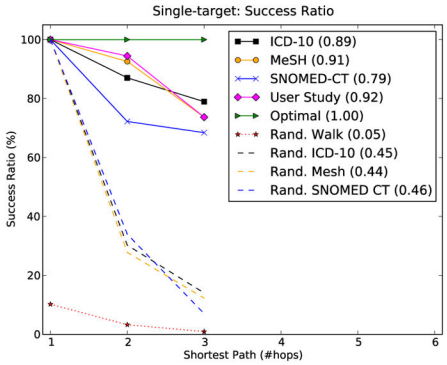
Disease articles commonly make use of an Infobox disease template, which offers fields for ontology codes. We used template fields in the Infoboxes to map Wikipedia articles to their ontology counterparts.

Fig. 4. Starting portals used in navigation simulations

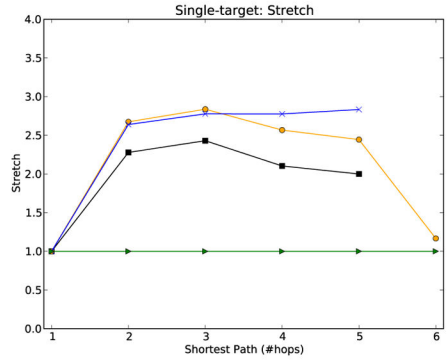
For ICD-10, MeSH and SNOMED CT we used a portal obtained by mapping navigation bar articles from WebMD.com to Wikipedia articles (subfigure a).



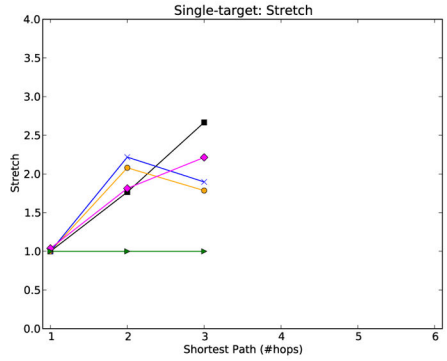
(a) Success Ratio for ICD-10/MeSH/SNOMED CT



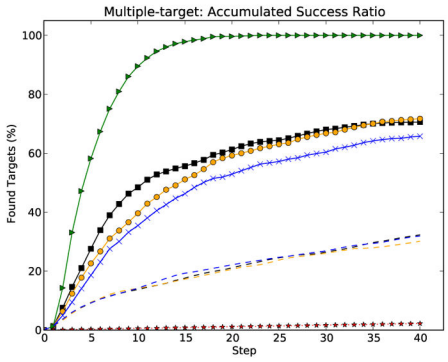
(b) Success Ratio for the User Study



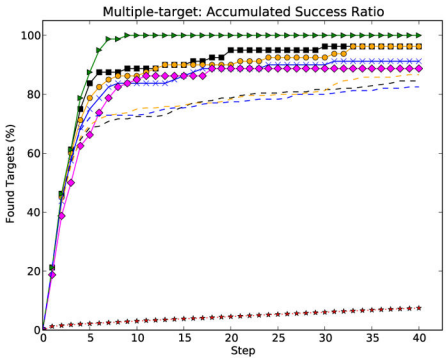
(c) Stretch for for ICD-10/MeSH/SNOMED CT



(d) Stretch for the User Study



(e) Accumulated Success Ratio for ICD-10/MeSH/SNOMED CT

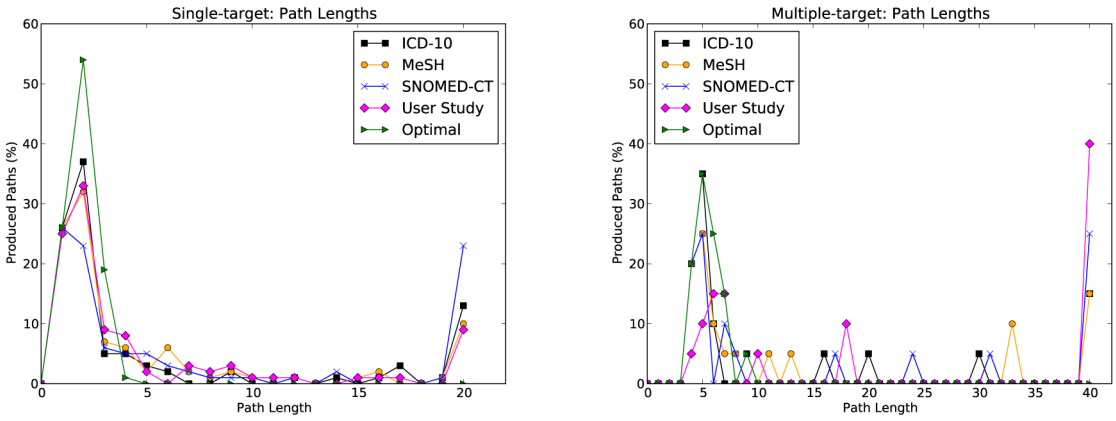


(f) Accumulated Success Ratio for the User Study

Fig. 5. Success ratio, stretch and accumulated success ratio for ICD-10/MeSH/SNOMED CT and the user study

The first column shows the results for ICD-10, MeSH and SNOMED CT, the second column the results for the user study. The rows show *stretch*, *success ratio* and *accumulated success ratio*, respectively. The legends in the first row are valid for the entire columns. The numbers in parentheses display the overall values for the success ratio. Note that the stretch plots do not include the random baselines, as this measure can only be usefully applied to compare simulations with a similar number of found paths. The figures show that the results produced by Ontology-based Decentralized Search are noticeably better than the results for

randomly generated ontologies. The figures also show that the results of Ontology-based Decentralized Search on a limited data set are in the range of the results produced by human test subjects.



(a) (b)

Fig. 6. Path lengths produced by the user study and the simulator

The figures show the resulting path lengths for the single-target (a) and multiple-target (b) search scenarios. Navigation was limited to 20 resp. 40 steps, hence the high number of paths for these lengths (i.e., not all targets were found). The path distributions for the random walk and the randomly generated ontologies were left out for reasons of clarity.

Table 1

Characteristics of the data sets used for our work

The tables displays statistics about the examined ontologies as well as the set of Wikipedia articles mapping to those articles.

		ICD-10	MeSH	SNOMED CT
Ontology	concepts	12,417	80,689	295,482
	top-level	22	16	19
	relations	12,416	112,463	440,408
	density	8.05×10^{-5}	1.73×10^{-5}	5.04×10^{-6}
	depth	4	14	16
	relation	is-a	is-a, part-of	is-a

Wikipedia	articles	1,593
	links	14,539
	density	5.73×10^{-3}

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 to ten articles.

Nausea-related	Stomach-related	Cough-related
Vomiting	Linitis plastica	Bronchitis
Nausea	Stomach cancer	Chronic bronchitis
Motion sickness	Gastritis	Acute bronchitis
Morning sickness	Atrophic gastritis	Cough
Drooling	Ménétrier's disease	Sputum
Hyperemesis gravidarum	Achlorhydria	
	Gastroparesis	
	Duodenal cancer	
	Gastric dumping syndrome	
	Stomach disease	

Table 3

Kullback-Leibler divergence for the path length distributions produced by the simulator and the user study. The table shows the KL divergence from the user study to the ontologies and the optimal and random solutions. The KL divergence measures the number of additional bits required to encode the original distribution, if another distribution is used in its place. The Randomly Generated Ontology column was computed using an average over the three randomly generated ontologies considered. The table shows, that the user study was more similar to the ontologies than to the base lines for the single-target scenario. The results closest to the user study are displayed in bold.

User Study	ICD-10	MeSH	SNOMED CT	Optimal	Randomly Generated Ontology	Random Walk
single-target	**0.12**	**0.08**	0.18	0.46	0.97	2.56
multiple-target	1.01	**0.74**	0.84	1.63	**0.55**	1.29

Table 4

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. The randomly generated ontology column was computed using an average over the three randomly generated ontologies considered.

		User Study	ICD-10	MeSH	SNOMED-CT	Optimal	Random Ontology	Random Walk
Found targets (Cosine Similiarity)	Single	1.00	0.93	**0.95**	0.89	**0.95**	0.78	0.72
	Multiple	1.00	**0.94**	**0.94**	0.91	**0.95**	0.90	0.67
First Hops (Cosine Similiarity)	Single	1.00	**0.89**	0.85	0.69	**0.88**	0.77	0.80
	Multiple	1.00	0.64	0.62	0.56	**0.68**	0.64	**0.71**
Back Button Uses (average per step)	Single	0.09	0.13	**0.11**	0.13	0.00	0.26	**0.07**
	Multiple	0.27	0.17	**0.18**	**0.18**	0.01	**0.21**	0.09
Home Button Uses (average per step)	Single	0.02	0.00	0.00	0.00	0.00	**0.01**	0.00
	Multiple	0.01	0.03	**0.02**	0.03	**0.00**	0.05	**0.00**

Semant Web. Author manuscript; available in PMC 2015 November 13.