

Using free link structure to calculate semantic relatedness

Sander Wubben

ILK Research Group Technical Report Series no. 08-01

Tilburg University

Faculty of Humanities

Department of Communication and Information Sciences

Tilburg, The Netherlands

July 2008

Abstract

In this thesis I present a new metric for the calculation of semantic relatedness. This metric uses the free link structure of conceptual networks to find shortest paths between concepts. I apply this metric to a conceptual network extracted from Wikipedia and a purpose build conceptual network: ConceptNet. These metrics are compared to existing metrics that use a hierarchical structure, such as WordNet. All metrics are tested on the Finkelstein-353 benchmark set, containing 353 wordpairs with a humanly assigned relatedness score. Finally, I demonstrate that a free link pathfinding measure based on Wikipedia is better for calculating semantic relatedness than existing WordNet measures.

Preface

“Tug on anything at all and you’ll find it connected to everything else in the universe.”

- John Muir

During the first half of the Master track I already thought a lot about possible topics for my Master’s thesis. I was very much intrigued by the vastness of Wikipedia and the possible ‘missing links’ that could be found by connecting articles. It was very pleasant that I got the opportunity to join the Dutch Common Sense team for the Battle of the Universities, because the Open Mind Commons project was very much in line with what I wanted to do in the first place.

Unfortunately the team did not make it to the next round, but the Dutch Open Mind Commons site is now a fact and I got to do the research I really wanted to be doing. The research process involved steering slightly away from common sense and more towards semantic relatedness. This is a very interesting area of research, and mining Wikipedia is another interesting activity that gets a lot of attention these days. I hope that by combining the two I have made a valuable contribution.

I would like to thank my supervisor, professor Antal van den Bosch. Despite being a very busy man, he managed to always make time somehow. Of course I also thank my girlfriend, Marianne, for listening to all my technobabble and never growing tired of it.

Contents

1	Introduction	1
1.1	Common sense	1
1.2	Semantic relatedness	2
1.3	Free link structure	2
1.4	Research question	3
1.4.1	Subquestions	3
1.5	Outline	3
2	Semantic Resources	4
2.1	WordNet	4
2.1.1	Structure	4
2.1.2	Organization	5
2.1.3	Quality	6
2.2	Wikipedia	6
2.2.1	Structure	7
2.2.2	Organization	8
2.2.3	Quality	9
2.3	ConceptNet	11
2.3.1	Structure	11
2.3.2	Organization	12
2.3.3	Quality	13

3	Lexical Semantic Relatedness Measures	14
3.1	Approaches using dictionaries or thesauri	14
3.2	Path-based measures	15
3.3	Information content based measures	16
3.4	Text overlap based measures	17
3.5	WikiRelate!	19
3.6	Wikipedia-based Explicit Semantic Analysis	20
4	Pathfinding In Graphs	22
4.1	Small-world networks	22
4.1.1	Scale-free networks	23
4.1.2	Preferential attachment	24
4.2	Pathfinding algorithms	24
4.2.1	Depth-first search	24
4.2.2	Breadth-first search	25
4.2.3	Weighted search	25
5	A New Measure: Free Link Pathfinding	27
5.1	Creating the network	27
5.1.1	Downloading the Wikipedia dump	27
5.1.2	Extracting the link structure	28
5.1.3	Indexing in- and outlinks	28
5.1.4	Using ConceptNet 3	29
5.1.5	Normalization	30
5.2	Implementing the search-algorithm	31
5.2.1	Calculating relatedness	31
5.2.2	Directed search	32
5.2.3	Undirected search	32
5.2.4	Weighted search	33

6 Experiments	34
6.1 The Finkelstein WordSimilarity-353 test collection	34
6.2 Correlation	35
6.3 WordNet experimental setup	36
6.4 Free link pathfinding experimental setup	36
6.4.1 Scale-freeness	36
6.4.2 Pathfinding	36
6.4.3 Modifying the network	36
7 Results	37
7.1 WordNet based measures	37
7.2 Free link pathfinding	39
7.2.1 Link distributions	39
7.2.2 Relatedness	40
7.2.3 Path lengths	41
8 Discussion	43
8.1 Network type	43
8.2 ConceptNet versus Wikipedia	44
8.3 Free link pathfinding versus other methods	44
8.4 Future research	45
References	46

Introduction

Why is it that computers can easily beat the best human chess-players or do billions of calculations on massive amounts of data, yet they lack any intelligence? Wouldn't it be great if a computer could actually understand what you tell it? Making machines understand language has been a focus of research ever since the day computers were created. The field that specifically focuses on automatically understanding texts is that of computational linguistics. This field of research has sprouted disciplines like machine translation, automatic text summarization, question answering, information retrieval, machine translation and so on. The problem in automating all these tasks is that humans always use their world knowledge when interpreting text, which a computer lacks. For example, in a conversation, humans will typically relate new information they receive to knowledge already in their possession, and from that infer assumptions and new knowledge. When we converse with others we expect our conversational partners to do the same. This is what tends to make our conversations interesting. To develop a machine that is able to do even a tiny part of the reasoning that humans do, it needs to have access to some kind of world knowledge. For humans, common sense is the most basic form of world knowledge.

1.1 Common sense

Researchers like Marvin Minsky and Doug Lenat have long argued that common sense constitutes the bottleneck for making intelligent machines ([Lenat and Guha, 1989](#)). Minsky describes how he worked for a couple of years on making a system that could understand the simple children's story:

"Mary was invited to Jack's party. She wondered if he would like a kite."

If you ask the question "Why did Mary wonder about a kite?", it is not hard to find the answer for any sensible human being: the party Mary was invited to was probably a birthday party, and if you go to a birthday party you bring a gift for the person that is celebrating his or her birthday. Jack is a boy and boys generally like things to play with like kites or balls. These things are all

knowledge we possess, and by inferring we can answer questions like this one. Minsky succeeded in making the computer understand this story, by putting these assertions in a database the system had access to. Unfortunately it failed on even a slightly different story. This led him to conclude that in order to have a computer that is able to reason in the way that we do, it would require a database with millions of assertions ([Minsky, 1986](#)).

1.2 Semantic relatedness

One step into gaining understanding into natural language is determining semantic relatedness, or its inverse, semantic distance between two concepts. Measures of semantic similarity are being used in applications such as text summarization and annotation, word sense disambiguation, information retrieval, automatic indexing, automatic correction of errors in a text and even automatic grading of essays.

It is important to make a distinction between the terms ‘semantic similarity’ and ‘semantic relatedness’. Semantic relatedness is a more general concept than semantic similarity. Semantically similar concepts are related due to their analogous nature: *bank* is the same as a financial institution. Similarity typically shows a synonymy relation. A lot of other relations are possible too: *car* and *engine* have a part-whole relation, *good* and *bad* have an antonym relation and intuitively we know that *snow* and *ski* also share a relation, but what kind of relation is sometimes hard to qualify. Concepts that are not considered semantically similar can very well be semantically related. The term semantic distance is somewhat ambiguous: it can mean the inverse of semantic similarity, but also the inverse of semantic relatedness.

In general computational linguistics applications benefit more from calculating relatedness rather than just similarity. When dealing with ambiguous words, the context is required to disambiguate. When we encounter *bank* and *money* in a text, we can disambiguate through relatedness. Yet most measures that are used nowadays calculate similarity instead of relatedness.

1.3 Free link structure

In an effort to provide machines with world knowledge, knowledge engineers have constructed numerous thesauri and ontologies that define in a formal way how certain concepts are defined and how they are related. This is typically done in a top-down way: the engineers have some vision about how the world works and they build their knowledge base from that vision. In recent years a lot of much hyped Web 2.0 applications have been brought forth. Some of these applications, like Wikipedia and ConceptNet, have made it possible to collect knowledge in a bottom-up way. An overview of how the world works emerges from all the contributions done by many users. These environments allow users to add knowledge in a way they choose themselves, and which is not necessarily pre-defined.

1.4 Research question

In this thesis I will investigate how free link structure can be used to calculate the relatedness between two given words. I will attempt to find a measure that is based on free link structure, and apply this measure to a purpose built conceptual network, namely ConceptNet, but also to a conceptual network extracted from the free link structure in Wikipedia. This measure will be compared with existing measures that use WordNet as their primary source. This leads to the following research question:

Is a measure based on free link structure valid for calculating semantic relatedness?

1.4.1 Subquestions

In order to calculate semantic relatedness using a free link structure, the measure that is developed needs to be scalable. Conceptual networks can become very large, and the amount of possible relations in such a network can easily get into the millions. Other relevant issues are which network is best for using the metric on, and if there are certain factors that improve the metric. Taking this into account, the following subquestions can be posed:

1. Are free link networks scale-free?
2. Is a network extracted from Wikipedia better than a purpose build network like ConceptNet?
3. Which factors benefit the computation of semantic relatedness?

These questions will be addressed throughout this thesis.

1.5 Outline

In this thesis, the second chapter will explore some of the different resources that are used for calculating semantic relatedness, namely WordNet, Wikipedia and ConceptNet. In the third chapter, different methods of calculating semantic relatedness that make use of lexical resources are investigated. The subject of chapter 4 is pathfinding in networks and in particular scale-free networks. The new free link measure is introduced and explained in chapter 5. Chapter 6 contains the description of the experiments conducted and in chapter 7 the results of these experiments are presented. Finally, in chapter 8 the conclusions and discussion can be found.

Semantic Resources

2.1 WordNet

WordNet is an electronic semantic lexicon for the English language. Its development started in 1985 under the direction of professor George A. Miller and is currently supervised by Dr. Christiane Fellbaum at Princeton University. WordNet can be regarded as an ontology for natural language terms. It attempts to model the lexical knowledge of a native speaker of English. Its design is based on psycholinguistic and computational theories of human lexical memory (Fellbaum, 1998). WordNet is widely used by researchers in among others the areas of computational linguistics and text analysis.

WordNet uses a differential theory of lexical semantics, meaning that representations are not on the level of individual words, but on the level of meanings of a word, called lexemes (Miller, 1995). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called synsets, each unambiguously expressing a certain concept. If a word has multiple meanings, they will be filed into different synsets. The result of this approach is that in WordNet a word is defined by its synonyms. Short general definitions called glosses are provided for each word and different relations link the synsets. The WordNet database and software tools have been released under a BSD style license and can be downloaded freely¹. The database can also be browsed online².

WordNet 3.0 contains over 155.000 words, grouped into over 117.000 synsets. Different WordNets have been developed and interlinked for several European languages in the EuroWordNet project (Vossen, 1998). These EuroWordNets are however not freely available.

2.1.1 Structure

The main structure of WordNet is that of a hierarchical network. Synsets can be related to other synsets in a variety of ways. The most common relation in WordNet is the hypernym/hyponym

¹<http://wordnet.princeton.edu/obtain>

²<http://wordnet.princeton.edu/perl/webwn>

(or IS-A) relation, which makes up for 80 percent of the total relations in WordNet. Concept X is a hypernym of Y if every Y is an X . This also makes Y a hyponym of X . So, the concept animal would be a hypernym of mammal and the concept mouse would be a hyponym of mammal.

This structure means that properties of concepts can be inferred: hyponyms inherit the properties from their hypernyms (or parents). If animals give birth, then mammals give birth as well, because they inherit this property. And if mammals suckle their young, mice suckle their young as well because they are mammals.

Another relation that is possible in WordNet is the holonym/meronym relation (or PART-OF). If Y is a part of X , X is the holonym of Y and Y the meronym of X . The concept computer is a holonym of CPU and keyboard may be a meronym of computer. Figure 2.1 shows an example of this structure.

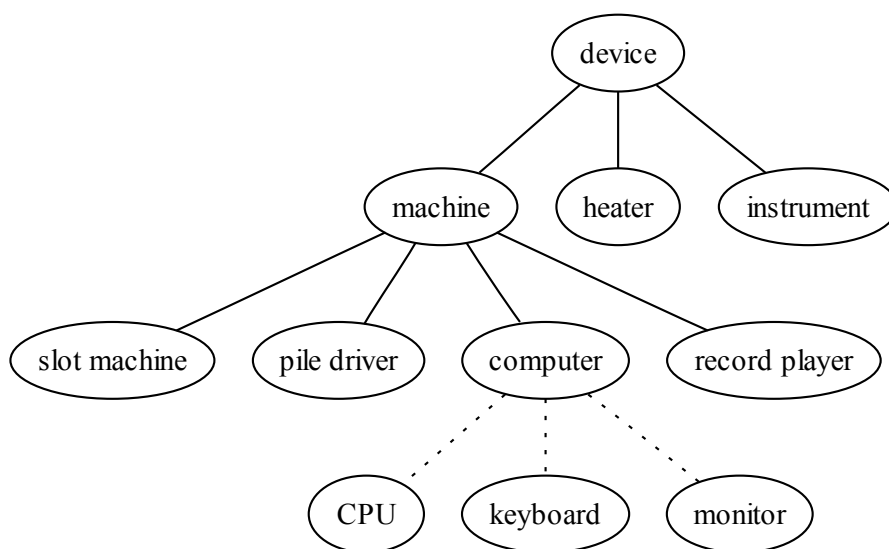


Figure 2.1: A fragment of the WordNet taxonomy. Continuous lines denote hypernymy relations, dotted lines holonymy relations

2.1.2 Organization

The largest part of the data in WordNet is generated by knowledge engineers in a top-down way. Sources that are used for data acquisition include monolingual dictionaries and lexical databases. WordNet's coverage is limited to the sources that are used. EuroWordNet additionally uses bilingual dictionaries to translate relations from one language to another. The Global WordNet Association³ was founded to discuss, share and connect WordNets for different languages in the world.

³<http://www.globalwordnet.org/>

2.1.3 Quality

Because of the free availability of WordNet and its clear documentation and software tools, researchers have used WordNet for a wide range of different applications in computational linguistics.

In information retrieval, query expansion can be used to increase the recall of a certain query. The results of early research in this area was not very good, especially when long queries were used (Voorhees, 1994). Later research expanded queries by adding parent and grandparent terms of the WordNet hierarchy to specific terms and children and grandchildren terms to abstract terms. In addition, all synonyms for a term were added to the query (Richardson and Smeaton, 1995). The precision of this system was however disappointing.

Mandala used WordNet as a tool for the automatic construction of thesauri, based on co-occurrence determined by automatic statistical identification of semantic relations, or on the predicate-argument association, in which the argument is constructed by identifying the most significant words of an environment (predicate) and those with which they relate (Mandala et al., 1998). Success was also achieved by Moldovan, who used WordNet for word sense disambiguation, increasing the precision of internet search by supplying a natural language interface (Moldovan and Mihalcea, 2000).

WordNet can be used to calculate semantic similarity, because of the information contained in the IS-A hierarchy. As is demonstrated in Figure 2.1, a computer and a record player can be thought of as being more alike than for example a computer and a tree, because *computer* and *record player* have a direct common ancestor in the IS-A hierarchy, while *computer* and *tree* do not.

2.2 Wikipedia

The encroaching rise of the Internet and the World Wide Web has enabled collaboration and co-operation on a global scale. The focus has shifted more and more from a ‘few to many’ to a ‘many to many’ perspective. The web encyclopedia Wikipedia⁴ is one of the best known examples of this process. Wikipedia is the world’s largest collaboratively edited source of encyclopaedic knowledge. Since its beginning in 2001 it has grown exponentially (Voss, 2005). There are Wikipedias available for more than 250 languages, of which 77 Wikipedias contain over 10,000 articles. Together these Wikipedias contain over 10 million articles, written by over 7 million contributors of whom 75,000 are regular editors. In the beginning of 2008 the English version contained over 2 million articles and received over 55 million visitors a month⁵.

Part of Wikipedia’s success is its implementation of wiki software. A wiki is a content management system that allows users to edit existing content of websites and create new content easily. This idea was introduced by Ward Cunningham, who started developing WikiWikiWeb in 1994 and it has turned out to work very well in the encyclopedic domain: everyone can extend or edit the

⁴<http://www.wikipedia.org/>

⁵<http://en.wikipedia.org/wiki/Wikipedia:Statistics>

encyclopedia. In their book *The Wiki Way: Quick Collaboration on the Web* (Leuf and Cunningham, 2001) described the wiki concept as follows:

- A wiki invites all users to edit any page or to create new pages within the wiki Web site, using only a plain-vanilla Web browser without any extra add-ons.
- Wiki promotes meaningful topic associations between different pages by making page link creation almost intuitively easy and showing whether an intended target page exists or not.
- A wiki is not a carefully crafted site for casual visitors. Instead it seeks to involve the visitor in an ongoing process of creation and collaboration that constantly changes the Web site landscape.

2.2.1 Structure

Everyone with access to the internet can write articles and edit most of Wikipedia's existing articles, as long as the article lives up to Wikipedia's editing policies. Contributors are endorsed to write neutral (NPOV: Neutral Point of View⁶) and obvious articles, citing sources, providing hyperlinks to other relevant articles and assigning categories to the article. A general Wikipedia page is structured as follows:

- Title
- First paragraph giving a general explanation or definition
- Overview of paragraphs
- Second and other paragraphs dealing with the subject matter
- Recommendations for further reading
- References
- Categories the article belongs to

Ambiguous terms are disambiguated on disambiguation pages (see Figure 2.2), listing all possible uses of the given term. Redirection pages make sure you end up on the right page, no matter how you write the word you are looking for (for example: *R2D2*, *r2d2*, *r2-d2* and *Artoo* all redirect to the page *R2-D2* about the Star Wars droid).

The main property of Wikipedia is that it is for the greatest part unstructured. On one hand editors are encouraged to supply their articles with categories. These categories can be part of larger categories, thus creating an ontology-like structure. On the other hand editors can link to any other page in Wikipedia, no matter if it is part of the same category, or any category for that matter.

⁶<http://en.wikipedia.org/wiki/Wikipedia:NPOV>

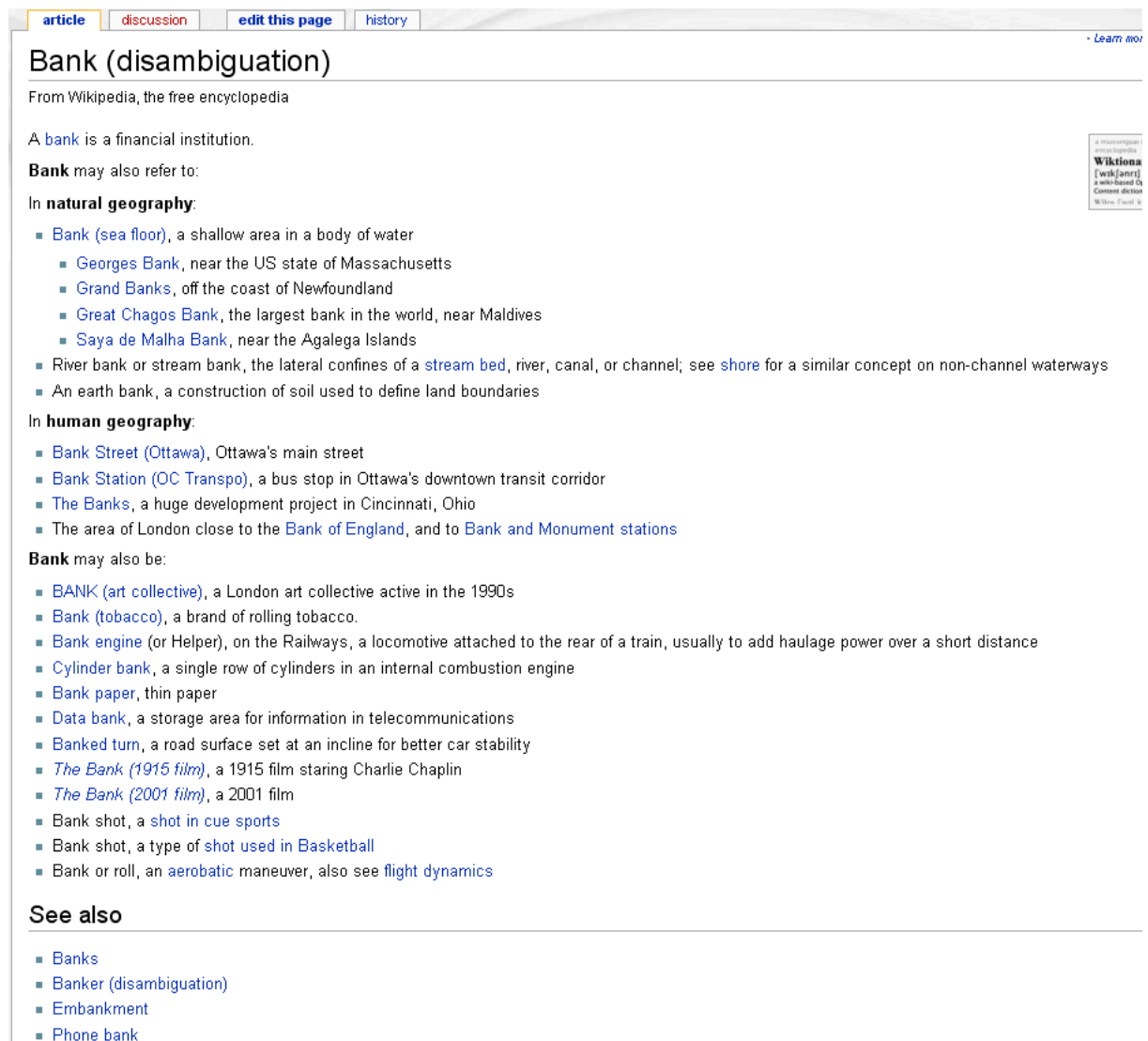


Figure 2.2: Disambiguation page for *Bank*

So although the category-graph is structured, the free links are not. An article can be assigned multiple categories, but the number of free links provided in an average article far exceeds the number of categories. This leads me to believe that Wikipedia is mainly an unstructured scale-free network, in the same way that the World Wide Web is. This bottom-up free architecture is radically different from the hierarchical top-down architecture as seen in WordNet. Scale-free networks will be explored further in chapter 4.

2.2.2 Organization

One of the key characteristics of Wikipedia (or for that matter any wiki), is that it lacks any top-down organization. As mentioned earlier, any visitor can not only read articles but also edit them or create new articles. This does not mean Wikipedia lacks organization. In editing Wikipedia, people assume different roles. There are readers, editors, administrators, recent changes check-

ers, stylists, policy makers, subject area experts, content maintainers, software developers, system operators and many more (Riehle, 2006). It can be concluded that people involved in maintaining Wikipedia largely organise themselves. Experts gather in different dedicated WikiProjects, to work on in-depth articles about one subject, others check lots of articles for errors or check whether the style lives up to Wikipedia's standards, while others are more involved in developing policies.

Analysis of the Dutch Wikipedia has shown that it can be considered an extreme form of self-management in regard to labour division. This bottom-up approach of self distribution of roles does not lead to chaos, but rather to an integrated and coherent data structure. Contributions roughly seem to follow a Pareto distribution (20 % of the contributors supplying 80 % of the content) (Spek, 2006).

2.2.3 Quality

One of the greatest criticisms on Wikipedia is that due to its nature of being an encyclopedia created and edited by anyone, it lacks authority and thus quality. When editing an article about for example *black holes*, edits made by a physicist do not weigh heavier than edits made by a layman. Another danger is people with double agendas. A lot of cases are known of people editing articles about themselves, one of the numerous examples is Adam Curry, a pod-cast pioneer who removed information about other pod-cast pioneers from Wikipedia. So, guidelines and policies alone cannot assure the quality of articles. A lot of this criticism came out of the corner of the traditionally edited paper encyclopedias, such as Britannica, who have seen their sales plummet since Wikipedia became a popular source of knowledge.

Their criticism does not appear to be totally justified, as research by Nature revealed. Nature compared the quality of Wikipedia with that of the Encyclopaedia Britannica. In their study, articles were chosen on a broad range of scientific disciplines and taken from both the Wikipedia and Britannica webpages. These were sent to relevant experts for a peer review. The reviewers were not told which article was from which website. The 42 returned reviews were used. Eight serious errors were found, four in each encyclopedia. Factual errors, omissions and misleading statements were more common. In Wikipedia 162 were found and in Britannica 123. So with regard to scientific articles the differences between Wikipedia and a 'proper' encyclopedia like Britannica are not that big (Giles, 2005). In June 2008 Britannica announced it is going to accept contributions from users, in a "collaborative but non-democratic" way.

The mechanism that seems to assure a great deal of quality in Wikipedia is the peer reviewed process: articles are under constant consideration of its viewers. Articles that need to be expanded or corrected receive banners stating so, inviting users to improve the articles. Of course, if an article is viewed often, errors in it will be corrected very quickly. Indeed, research has shown that articles with many edits have a higher quality than articles that have been edited less frequently. Because popular articles receive more views, they are edited more heavily as well. This means popular articles are in general of higher quality than less popular articles (Wilkinson and Huberman, 2007). In addition to that, articles are not only checked by humans, but also by bots. Bots may check

articles for quality, but also edit out commonly made mistakes.

One feature that takes full advantage of the peer-reviewed process is the existence of the earlier mentioned WikiProject pages. Experts in one specific area can organise themselves in these projects and work collaboratively on developing high quality content for the articles in their interest areas.

Another feature that helps assure Wikipedia's quality is the discussion page that is linked to each entry. When editing controversial articles like those about *Global Warming* or *Palestina*, on these discussion pages users can argue with each other, providing arguments and sources for their viewpoints. This prevents edit wars, where groups of users with different viewpoints keep editing out any changes made by the opposing group.

A third feature of Wikipedia that helps quality is the process of featuring high quality articles on the Wikipedia frontpage. These so called Featured Articles (FA) are elected through well established and visible processes (Figure 2.3). The requirements for promoting an article to FA status have increased dramatically over the years. Nowadays, over 200 of the early FAs have been demoted because they do not meet the current FA criteria. The first step towards the FA status for an article is to become a Featured Article Candidate (FAC). In order to become a FAC, someone needs to nominate the article. These nominations are public and when editors raise objections the nominators are expected to address these objections by editing the article. In order to be promoted from FAC to FA, consensus must be reached. Anyone can cast a vote for or against promotion, provided it is backed by explicit reasoning. The FA director finally decides when consensus is reached. Different tools are used to smooth the process, such as citation checking scripts and workflow templates (Viégas et al., 2007).

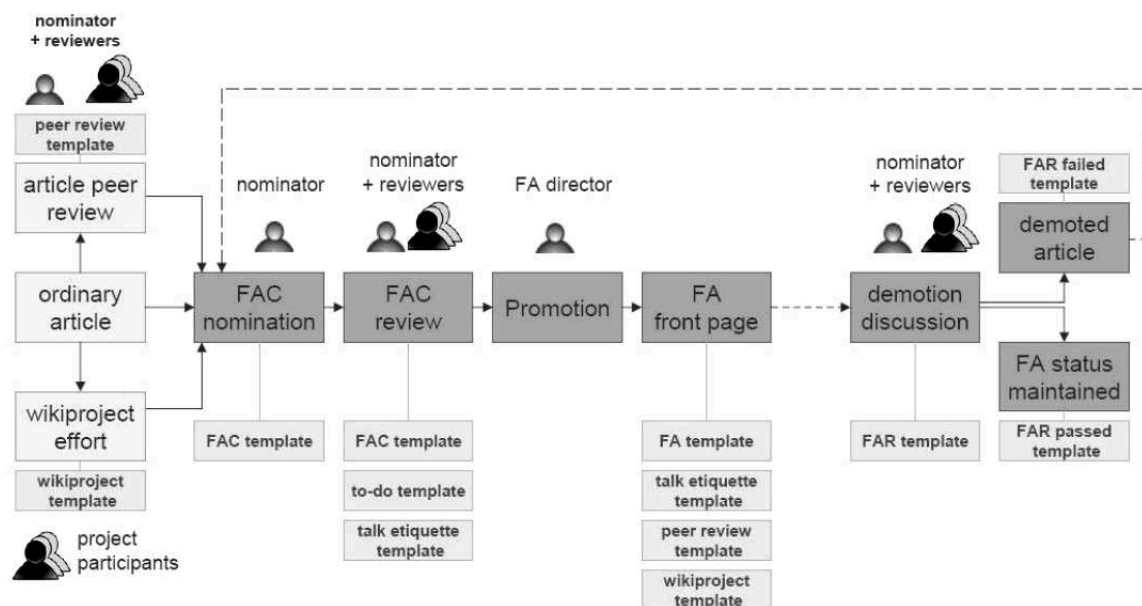


Figure 2.3: Overview of the Featured Article promotion and demotion processes, taken from (Viégas et al., 2007)

2.3 ConceptNet

ConceptNet⁷ is a freely available, machine-usable common sense resource. ConceptNet 3 presently consists of over 250,000 elements of common sense knowledge, in the form of semi-structured fragments of natural language. The creation of ConceptNet was inspired by the large amount of common sense concepts and relations in Cyc (Lenat, 1995), and by the ease-of-use of WordNet. The representation of a semantic network of WordNet was used, but augmented in several ways (Liu and Singh, 2004), (Havasi et al., 2007). Nodes in ConceptNet represent concepts and edges represent relations (Figure 2.4).

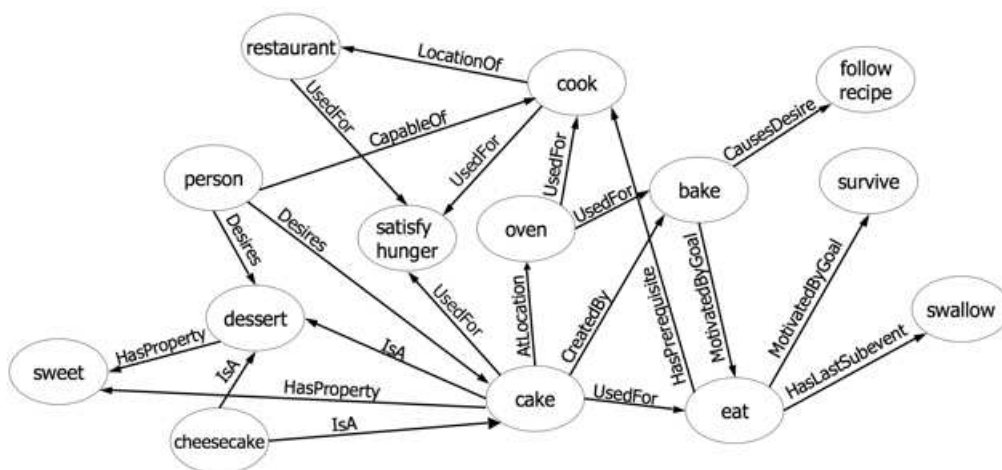


Figure 2.4: An example of the ConceptNet 3 structure

2.3.1 Structure

Nodes in ConceptNet can, in addition to noun phrases which are typically found in WordNet (like for example *food* and *junk food*), also encompass higher-order compound concepts in the form of verb phrases. These verb phrases match an action verb with one or two direct or indirect arguments (for example: *buy food, drive a car*). Knowledge about a greater deal of real world concepts can be represented in the semantic network in this way. The downside to this approach is that ConceptNet does not distinguish between word senses.

A second difference with WordNet is that ConceptNet extends WordNet's small ontology of mainly taxonomic semantical relations to include a richer set of possible predicates that express the relations between concepts. These predicates represent the edges and are, just like the nodes, presented in natural language. Some examples hereof are displayed in Table 2.1. In addition to these specific relation types, a relation can also be unspecified, such as *ConceptuallyRelatedTo*, which says that two concepts are related, but the nature of the relationship cannot be determined.

⁷<http://conceptnet.media.mit.edu/>

Relation	Example sentence patterns
IsA	NP is a kind of NP.
MadeOf	NP is made of NP.
UsedFor	NP is used for VP.
CapableOf	NP can VP.
DesireOf	NP wants to VP.
CreatedBy	You make NP by VP.
InstanceOf	An example of NP is NP.
PartOf	NP is part of NP.
EffectOf	The effect of VP is NP VP.

Table 2.1: Some of the specific predicates in ConceptNet 3, along with an example of a sentence pattern that produces each predicate.

The final difference between ConceptNet and WordNet is the nature of the information present. ConceptNet contains more informal knowledge of practical value. In WordNet a *dog* is a *canine* and a *cat* is a *feline*, and both are *carnivores*, *placental animals*, etcetera. But WordNet does not contain the knowledge that both cats and dogs are pets. The taxonomical information about cats and dogs is less likely to enter into ConceptNet than practical information about these concepts. In addition, ConceptNet also contains a lot of defeasible information: facts that are often true, but not necessarily always, like: *EffectOf*(‘fall off bicycle’, ‘get hurt’). In our everyday lives we deal with these kind of fuzzy truths all the time.

2.3.2 Organization

While CYC and WordNet take their knowledge from the inputs of knowledge engineers, ConceptNet is in the aspect of knowledge acquisition more like Wikipedia: it takes its knowledge from contributors on the Internet. Users can add assertions using predefined formats, such as: You are likely to find *A* in *B*. The user can fill in anything he wants in the *A* and *B* spots. Another important feature of the Open Mind Commons (OMCS) website⁸ is the ability to provide feedback. Users can rate each others assertions positively or negatively, influencing the score of those assertions. The higher the score, the more reliable a statement is considered. Every time a user adds an assertion that is already present in the database or rates an assertion positively, its score is increased by one.

OMCS also asks feedback in a different way. If concept *A* and concept *B* appear in corresponding positions in many similar predicates, they can be considered similar to each other. If concept *A* appears in a predicate that *B* does not appear in, OMCS can infer that the same predicate might be true for *B*. Because of the natural language interface to ConceptNet, OMCS can return this inferred predicate to a user and ask if it makes sense. It also asks users to fill in the blanks on concepts that do not have enough predicates. Analogous to other concepts it will ask question that the user can evaluate as true or false (Figure 2.5).

























⁸<http://commons.media.mit.edu/en/>

Concept: oven

Similar concepts

kitchen, refrigerate, table, restaurant, cupboard, supermarket, grocery store, cabinet, plate, Fridge

Current knowledge

→ <u>an oven</u> is used for <u>cooking food</u>	by  Visionsoftkaos	Score: 6	 
→ You can use <u>an oven</u> to <u>bake</u>	by  Laserjoy	Score: 6	 
→ You are likely to find <u>an oven</u> in <u>the kitchen</u> .	by  egamma	Score: 6	 
→ Something you find in <u>the oven</u> is <u>cakes</u>	by  AnotherBackAche	Score: 5	 
→ Something you find in <u>the oven</u> is <u>heating elements</u>	by  Visionsoftkaos	Score: 5	 
→ Something you find in <u>the oven</u> is <u>racks</u>	by  Visionsoftkaos	Score: 5	 
→ Something you find in <u>the oven</u> is <u>bread</u>	by  Visionsoftkaos	Score: 4	 
→ <u>baking a cake</u> requires <u>an oven</u> .	by  papayaimae	Score: 4	 

Page 1 of 13 | [Next](#) | [Last](#) (102 total)

Open Mind wants to know...

You are likely to find in

You are likely to find in

is for

Figure 2.5: The Open Mind page for 'Oven', giving users the ability to provide feedback in two ways: judging assertions and judging or altering inferences from the system

2.3.3 Quality

A study by (Singh et al., 2002) was carried out to determine the quality of the OMCS 1 database. About 3000 standard items were randomly selected and judged manually. Of these items, 12.3 % was marked as garbage and thus unusable. The remaining items were judged on generality, truth, neutrality and sense on a scale from 1 (worst) to 5 (best). Generality scored on average just above 3 and all the other attributes scored over 4. The items were also judged on age level. 84 % of the items was judged to be on grade- or high school level, indicating that most of the database indeed consists of facts most people know, and thus can be considered common sense.

Lexical Semantic Relatedness Measures

There are two general approaches to measuring semantic relatedness: one approach that uses lexical resources for measuring semantic relatedness and one approach that uses distributional statistics of words in a corpus to measure semantic similarity. I will focus mainly on the first approach: lexical semantic relatedness. Lexical semantic relatedness measures take a lexical resource and transform this resource into a network or graph and compute semantic relatedness by using the paths that exist in the generated graph.

3.1 Approaches using dictionaries or thesauri

Kozima and Furugori used the Longman Dictionary of Contemporary English as a lexical resource and translated it into a semantic network. Every headword in the dictionary was turned into a node, and each node was connected to other nodes of words that occurred in the definition of the headword. Similarity between words is computed by spreading activation on the semantic network. Each word is represented by all the words in its definition ([Kozima and Furugori, 1993](#)).

A thesaurus is similar to a dictionary, but contains relations such as synonyms and antonyms. Roget's Thesaurus was the first of thesauri. It was compiled in 1805 by Dr. Peter Mark Roget, and published in 1852. It has been updated ever since and now contains over 250,000 words, starting with 15,000 words back in 1852. Unlike in a dictionary, entries in Roget's Thesaurus are listed conceptually rather than alphabetically and there are no definitions for words. Roget's Thesaurus is structured into six primary classes. Each class is composed of multiple divisions and each division is divided into sections. Semantically related words become clustered in categories in one of the many branches of this system. Polysymy can be solved by looking at the other words in the cluster and by their index entry. The index entry for each word contains category numbers and labels. Categories can contain pointers to other categories.

Using Roget's Thesaurus, (Morris and Hirst, 1991) identified five types of semantic relations between words. If any of the following conditions are met two words are considered similar.

1. Both words have a category in common in their index entries.
2. One word has a category in its index entry that contains a pointer to a category of the other word.
3. One word is either a label in the other word's index entry or is in a category of the other word.
4. Both words are contained in the same subcategory.
5. Both words have categories in their index entries that point to a common category.

3.2 Path-based measures

The most basic way of computing semantic similarity between two concepts c_1 and c_2 is measuring the distance in a semantic network such as WordNet between c_1 and c_2 . This can be achieved by finding the paths from each sense of c_1 to each sense of c_2 , and then taking the shortest. This results in the semantic distance. The semantic distance is inversed to get the semantic similarity. Computing the path length between c_1 and c_2 can be done using the formula

$$sim_{path}(c_1, c_2) = max[\frac{1}{Np}]$$

where Np is the number of nodes in path p (see Figure 3.2(a)). This simple representation is based on the notion that all distances between nodes are equal. This is typically not the case. Resnik points out that the basic path length measure will suffer from the great differences in depth found in different parts of the taxonomy (Resnik, 1995). This is due to the fact that some classes are far more specific than others.

To overcome this problem, Leacock and Chodorow proposed a normalized path-length measure (lch) which also considers the depth of the taxonomy that is used:

$$sim_{lch}(c_1, c_2) = -\log \frac{length(c_1; c_2)}{2D}$$

where $length(c_1; c_2)$ is the number of nodes along the shortest path between the two nodes (the basic path length), and D is the maximum depth of the taxonomy (Leacock et al., 1998).

Wu and Palmer presented in a paper on translating verbs from English to Mandarin Chinese a scaled measure (wup) which measures what they call conceptual similarity and takes into account the depth of the nodes together with the depth of their most-specific common subsumer (Wu and Palmer, 1994). This hypernym is also known as *the lowest superordinate* (LSO).

$$sim_{wup}(c_1, c_2) = -\log \frac{depth(LSO(c_1; c_2))}{depth(c_1) + depth(c_2)}$$

3.3 Information content based measures

The method Resnik uses to solve the problem of determining the importance of a category hinges on the intuition that one criterion of similarity between two concepts is the extent to which they have common attributes. In a IS-A taxonomy this can be determined by inspecting the relative position of the LSO. Instead of using path length, Resnik uses the most informative class to compute similarity, so the structure of the semantic network is only used to find the LSO. A class consists of all the synonyms found at the LSO and all the synonyms of its hyponyms. To find the informativeness of the classes Resnik gathered their frequencies from the one-million-word Brown Corpus of American English. For example: to compute the frequency of the class *money* all occurrences of the word *money* and all its defined synonyms are counted, as well as the occurrences of all hyponyms such as *nickel* and *dime* and their defined synonyms. The frequencies are then adjusted to take into account the number of classes a word belongs to. Using information theory, the probability of the classes can then be determined. The similarity between word c_1 and word c_2 is determined by the most informative (and thus least probable) class they belong to:

$$sim_{res}(c_1, c_2) = -\log p(ISO(c_1, c_2))$$

with p being the probability of the class both words belong to. The probability p can be calculated by counting the frequencies in a corpus:

$$p(c) = \frac{\sum_{w \in W(c)} count(w)}{N}$$

where $W(c)$ is the set of words that are subsumed by concept c and N is the total number of words that are present in the corpus and the taxonomy (Resnik, 1995). Figure 3.1 demonstrates that

$$sim_{res}(dime, creditcard) = sim_{res}(money, credit)$$

because both pairs share the same LSO, while

$$sim_{path}(dime, creditcard) < sim_{path}(money, credit)$$

because the path between dime and credit card is longer than the path between money and credit. Additionally, the similarity will decrease as the LSO is situated higher in the taxonomy, because then it becomes more abstract and as a result more probable. If the LSO is the top node, its probability will become 1 and thus the similarity $-\log(1) = 0$ (Budanitsky and Hirst, 2006).

The problem with many semantic similarity measures is that they are specifically tailored for one domain. To overcome this problem, (Lin, 1998) attempted at a similarity measure that would be universally applicable and theoretically justified. He based his measure on the three intuitions that:

1. The similarity between objects A and B is related to their commonality; the more commonality they share, the more similar they are.
2. The similarity between A and B is related to the differences between them, the more differences they have, the less similar they are.

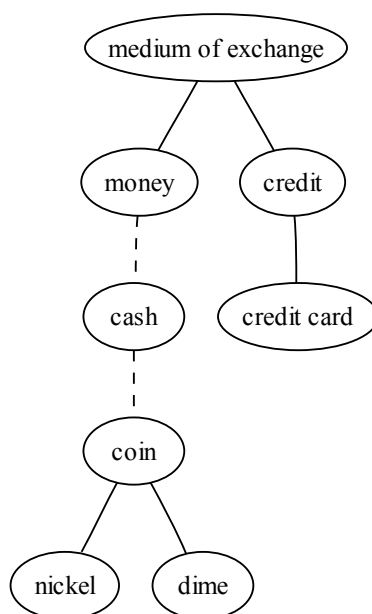


Figure 3.1: An example of the WordNet taxonomy, showing lowest superordinates of nickel and dime (coin) and of nickel and credit card (medium of exchange). Dashed lines indicate that some intervening nodes have been left out. Adapted from (Resnik, 1995).

3. The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

Both the method used by Lin (*lin*) and the method used by (Jiang and Conrath, 1997) (*jcn*) augment the information content of the LSO of two concepts with the sum of the information content of the individual concepts. The difference between the two is that *lin* scales the information content of the LSO by this sum, while *jcn* subtracts the information content of the LSO from this sum, and takes the inverse of this number to convert it from a distance to a similarity measure.

3.4 Text overlap based measures

For word sense disambiguation, Lesk constructed an algorithm based on the idea that related words are often defined using the same words. Given a word to disambiguate, the original Lesk algorithm compares the definition text of each sense of that word (the glosses of that word) from a dictionary with the glosses of every other word in the sentence. The sense whose gloss shows most overlap with the glosses of the other words will then then picked (Lesk, 1986). Overlap is calculated by counting the number of content words in common.

The working of the algorithm can be demonstrated by considering the words *pine cone*. The algorithm, using the Oxford Advanced Learner’s Dictionary, finds the following two senses for *pine*:

1. Kind of **evergreen tree** with needle shaped leaves.
2. Waste away through sorrow or illness.

For *cone* it finds three senses:

1. Solid body which narrows to a point.
2. Something of this shape whether solid or hollow.
3. Fruit of certain **evergreen tree**.

The glosses of sense one for *pine* and sense three for *cone* show the largest overlap, so these senses are picked for *pine cone*. A drawback of Lesk's approach, is that dictionary glosses tend to be fairly short and thus do not provide sufficient vocabulary to make subtle distinctions in degrees of relatedness. (Banerjee and Pedersen, 2003) expanded Lesk's approach to include the glosses of other concepts to which the senses of the words under considerations are related according to a given concept hierarchy such as WordNet. The advantage is that the glosses of these words can be expanded by the words in the hierarchy and also that relations that are not explicit in the hierarchy can implicitly be observed through the gloss overlap. For example, in WordNet, *car* and *tyre* do not share a relationship, while their glosses show a large degree of overlap. In addition to that, *vehicle* and *car* do share an IS-A relationship, so the gloss of *car* can be extended by the gloss of *vehicle*. Figure 3.2(c) shows another example.

(Patwardhan and Pedersen, 2006) expanded this approach with their gloss vector by applying second order co-occurrence vectors on the WordNet glosses. They based their research on the assumption that vectors built from the contexts of words are useful representations of word meanings. This was demonstrated earlier by (Schutze, 1998).

For example, *car* and *mechanic* are likely first order co-occurrences since they commonly occur together. A first order context vector for a given word simply indicates all the first order co-occurrences of that word as found in a corpus, such as the collection WordNet glosses. Because the Gloss Vector measure is based on second order co-occurrences, it includes the contexts of *mechanic* and *car* as well. *Mechanic* and *police* are second order co-occurrences since they are both first order co-occurrences of *car*. A spatial representation is in Figure 3.3.

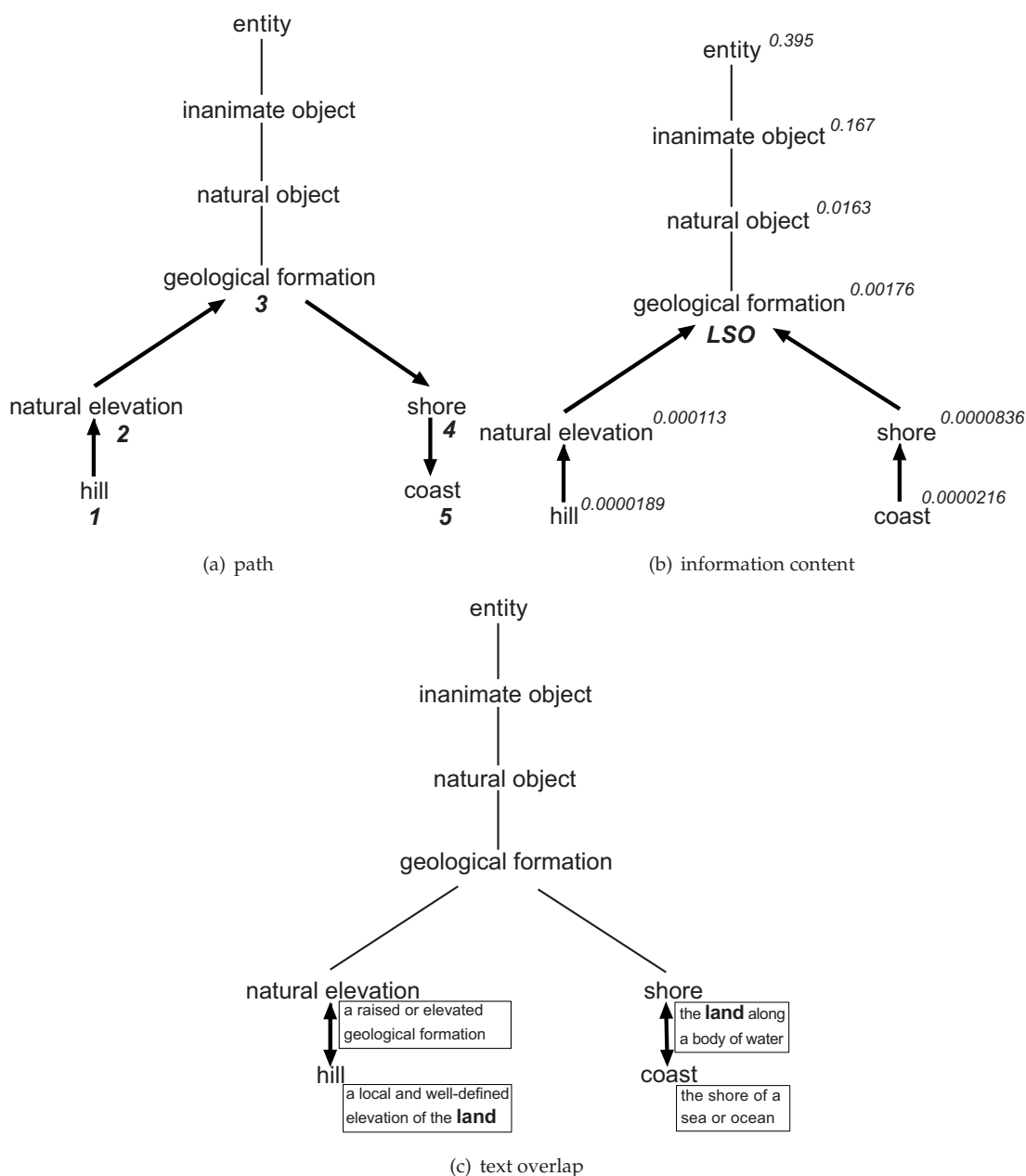


Figure 3.2: Examples of calculating $\text{sim}(\text{hill}, \text{coast})$ in WordNet using three different approaches

3.5 WikiRelate!

The semantic relatedness measures originally developed for WordNet were applied to Wikipedia by (Strube and Ponzetto, 2006). Users can assign categories to their Wikipedia articles. These categories can be part of larger categories, thus forming a hierarchical graph. This category graph was used to construct a semantic network. For every word pair under consideration, WikiRelate! first retrieves the Wikipedia pages the words refer to. These pages are then hooked to the cate-

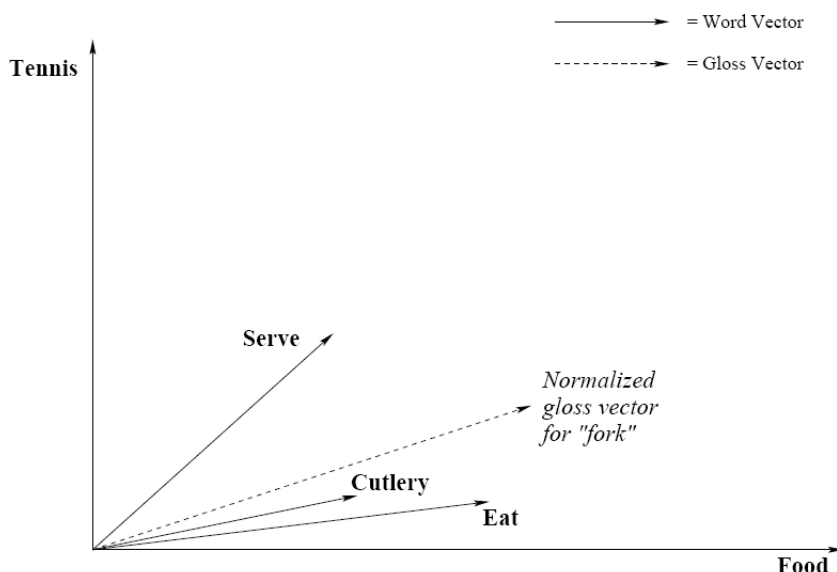


Figure 3.3: First Order Context Vectors and a Gloss Vector

gory tree by extracting the categories each page belongs to. Finally the path can be found between the extracted categories along the Wikipedia category graph. In the case of ambiguous concepts, Strube and Ponzetto try to let the queries disambiguate themselves. If a disambiguation page is hit, all links from that page are used as a lexical association list. Words between parentheses are split and added to the list. If overlap is found between the lexical association list of word *A* and word *B* (or overlap between the two lists, if both words need to be disambiguated), that link is followed and the target becomes the concept to use. If no overlap is found the first redirect on the disambiguation page is used.

If the pair under consideration is *King* and *Rook*, first both concepts need to be disambiguated. *King* points to a disambiguation page linking to among others *Monarch*, *King Kong* and *King (chess)* while *Rook* leads to a disambiguation page redirecting to among others *Rook (chess)*, *Rook (bird)*, *Rook (rocket)*. *Chess* shows up in both lists, meaning that those specific redirects are used for both concepts (*King (chess)* and *Rook (chess)*).

The Resnik, Wu & Palmer and Leacock & Chodorow algorithms all show a large increase in correlation with human judgements on the Finkelstein-353 dataset when using the WikiRelate! system over WordNet 2.0. This clearly shows that the larger coverage of Wikipedia is of great benefit to semantic relatedness measures.

3.6 Wikipedia-based Explicit Semantic Analysis

Gabrilovich and Markovitch recognized the need to augment texts with common sense knowledge to compute semantic relatedness. They proposed Explicit Semantic Analysis (ESA), a method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia.

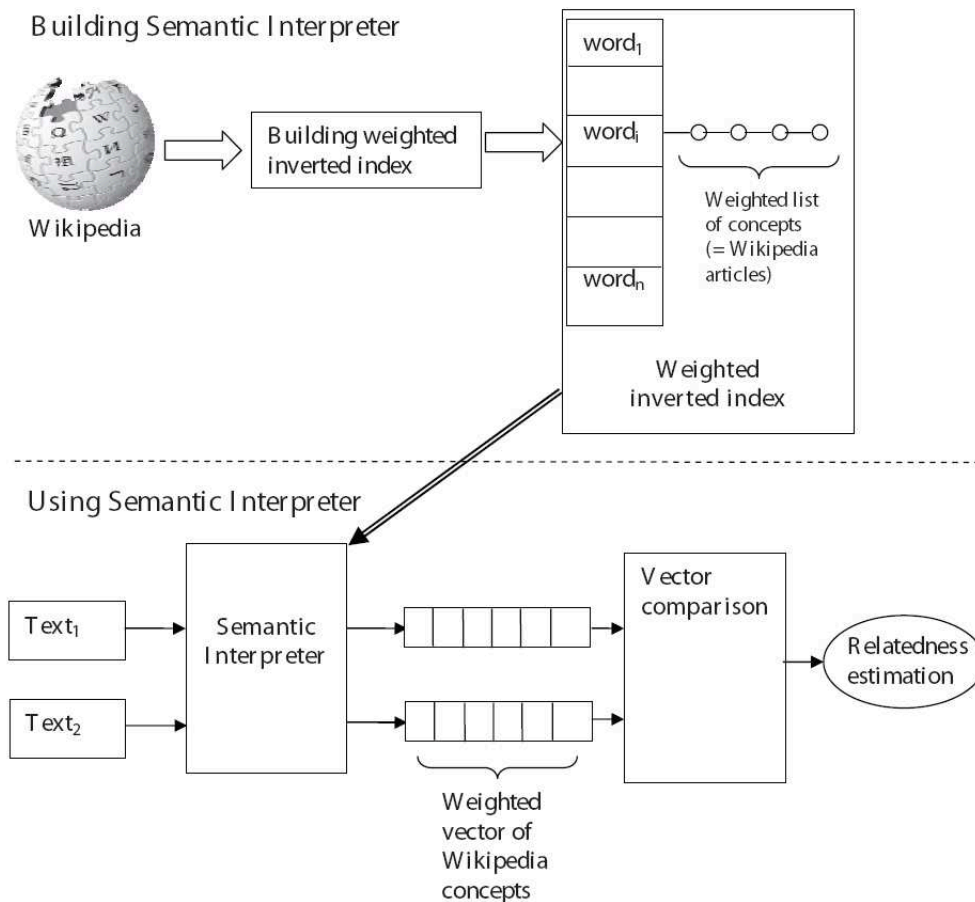


Figure 3.4: Schematic of the Wikipedia-based Explicit Semantic Analysis.

They used machine learning techniques to build a semantic interpreter that maps fragments of natural language text into a weighted sequence of Wikipedia concepts that are ranked by their relevance to the text. This means that texts are explicitly represented by weighted vectors of concepts (interpretation vectors) (Gabrilovich and Markovitch, 2007).

The meaning of a text fragment is thus interpreted in terms of its affinity with a range of Wikipedia concepts. Semantic relatedness is then computed by comparing the vectors of the texts in the space defined by the Wikipedia concepts (Figure 3.4). This can be done using conventional methods such as the cosine metric. The representation of texts is explicit in that way that representations are in natural language concepts present in human cognition, as opposed to Latent Semantic Analysis (LSA), which uses abstract latent semantics.

Pathfinding In Graphs

Pathfinding is the process of plotting the shortest route from point A to point B. In hierarchical graphs this is an easy task: just go up in the hierarchy until the lowest superordinate of both A and B is found. In small-world networks the task is more challenging.

4.1 Small-world networks

A small-world network is a certain type of graph in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of steps through the graph. The small-world phenomenon became known by an experiment in the science of social networks by Stanley Milgram of Harvard University. Milgram noticed that people's friendship circles are often highly clustered. These clusters can be linked by people who are members of various clusters, allowing even large communities to be quite cohesive. To test this idea, he concocted an experiment to see how well connected the world really was. In his experiment, performed in 1967, Milgram randomly chose a stockbroker near Boston and 160 residents of a small town near Omaha, Nebraska. He sent the residents of the town each a package and instructions to send the packages by mail to the Boston stockbroker identified only by his name, occupation and rough location. They were not allowed to look him up in a telephone book and could only send the package to the stockbroker himself or to someone in their social network they knew on a first-name basis and whom they thought would most likely get the package further on the way to the stockbroker. Milgram found on average it took only six intermediaries to link the two people (Milgram, 1967). Later this concept of 'six degrees of separation' was popularized by John Guare's play by the same name and the game "Six Degrees of Kevin Bacon" developed at the University of Virginia¹, which links actors by co-occurrence in movies to Kevin Bacon. This game was inspired by the Erdos number, a similar metric, but instead of actors it links scientists who collaborated on writing articles to mathematician Paul Erdos.

¹<http://oracleofbacon.org/>

4.1.1 Scale-free networks

Scale-free networks are a special kind of small-world network. In a scale-free network most nodes have a low connectivity. There are however a number of nodes with a very high degree of connectivity. This behaviour was discovered by (Barabasi and Albert, 1999), who crawled the web to map its connectedness. They discovered that the web was not connected randomly, but that certain nodes had many more connections than average. These nodes act as hubs: they connect the various parts of the network (see Figure 4.1). The structure and dynamics of these networks are independent of the scale of the network, hence the name scale-free. The heavily tailed and right skewed distribution of the degree of connectivity of the nodes follows a power law, which is defined by

$$P(k) \sim k^{-\gamma}$$

where the probability $P(k)$ that a node in the network is connected to k other nodes is roughly proportional to $k^{-\gamma}$. The coefficient γ varies approximately from 2 to 3 for most real-world networks. Figure 4.2 compares this distribution to the bell-shaped distribution of random networks. The existence of these well connected nodes ensures that path-lengths do not grow significantly when the network grows. This means the network can grow endlessly without losing its usability. Another benefit of scale-free networks is that it is very resistant to failure. A lot of nodes can be turned off, without the network losing its connectivity. Only when a number of highly connected nodes is targeted does the network start to fail.

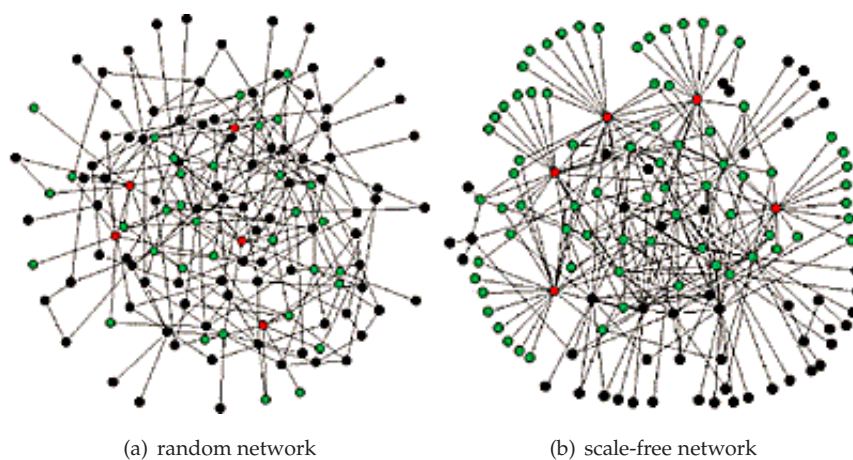


Figure 4.1: A random and a scale-free graph. Highly connected nodes are colored red.

Interestingly, many real world phenomena have scale-free characteristics: social networks, computer networks, semantic networks, the spreading of viruses (both the real world and the computer variants), public transport (airports are obvious examples of hubs), but also cellular metabolism and even the working of brain functions (Eguíluz et al., 2005).

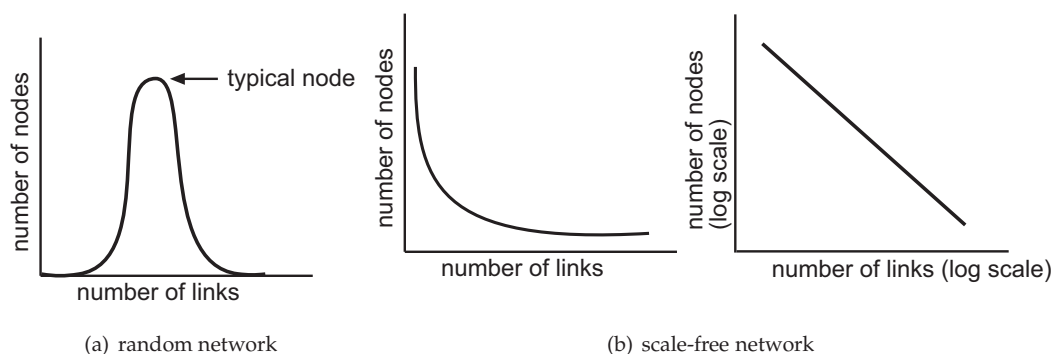


Figure 4.2: Connectedness distributions of the nodes in a random and scale-free graph.

4.1.2 Preferential attachment

So how can the emergence of scale-free networks be explained? The answer is a sort of ‘the rich get richer’ process. Well connected nodes in a network are more likely to receive new connections than poorly connected nodes. When for example a new person enters a social network, the chance of him getting acquainted with well-connected people in that network is relatively high, because he has easy access to those persons; through their relations. The same is true for the World Wide Web: when a new webpage is created, the chance is high it will link to well established sites, because the maker of a new website is more likely to know about those well-connected sites.

4.2 Pathfinding algorithms

A great benefit of scale-free networks is that path lengths do not grow to endless lengths when the network grows. This means these networks can be traversed no matter how large they grow. Of course, to do this automatically, efficient algorithms are needed. Because of the small wordliness of such a network, all nodes are close to each other, which means the number of possible paths rises extremely fast for each extra step, specially in large networks.

4.2.1 Depth-first search

A depth-first search (DFS) algorithm takes a certain node in the graph as its root and explores as far as possible along each branch originating from the rootnode as possible. The algorithm traverses the graph deeper and deeper, until the goal node is found, or a dead end is encountered. Then it backtracks to the previous node that was not explored fully yet, following that one as deep as it can. These nodes are kept in a stack, meaning that nodes that go in last will go out first (LIFO). Of course, a list of already explored nodes needs to be maintained as well, to prevent loops.

DFS lends itself well to heuristic methods of choosing a branch that is likely to be the best. Space complexity of DFS is much lower than for example breadth-first search. Its time complexity is proportional to the number of vertices plus the number of edges in the graph that is traversed, in

big O notation: $\mathcal{O}(|V| + |E|)$. When searching in large graphs, the list of nodes that have already been visited by DFS can grow extremely large, as do the paths that DFS takes. This can be solved by limiting the depth of the tree. This method is called iterative deepening depth-first search.

DFS can be used to find connected components, for topological sorting or to solve problems that have only one solution. In that regard it is very similar to the classic method of finding a path through a maze. If the graph contains cyclic paths DFS will not always find the shortest path.

4.2.2 Breadth-first search

A breadth-first (BFS) algorithm also begins in a root node, but unlike DFS it explores all of that node's neighbouring nodes first. For each node it has encountered, it then explores all its yet unexplored neighbouring nodes. In this way the algorithm builds an evergrowing front of nodes to explore, until it encounters the target node. Nodes that need to be explored are stored in a queue: nodes that go in first, go out first (FIFO). This means that the network is traversed layer by layer.

If the graph depth is d and the branching factor of the nodes in the graphs is b , then in big O notation the space and time complexity of BFS asymptotically approaches $\mathcal{O}(b^d)$. However, as we have seen earlier, in small-world networks d is usually not higher than 6.

BFS will always find a path from the source- to the target node and it will always find the shortest path possible, meaning BFS is complete.

Bidirectional breadth-first search

Breadth first search can be made more time and space efficient by dividing the task in two parts. Instead of only expanding the source node until the target node is hit, bidirectional search expands both the target and source nodes, until the two fronts 'hit' each other. This approach shrinks space and time complexity to $\mathcal{O}(b^{\frac{d}{2}} + b^{\frac{d}{2}})$.

4.2.3 Weighted search

The algorithms described above all assume that the edges between nodes are all equal in length. When lengths or costs are assigned to edges the pathfinding task becomes different: it is no longer sufficient to find the path with the shortest amount of edges. Instead, now the task is to find the path with minimal costs. This is a far more complex problem than breadth-first search, because when a path is found it is not necessarily the shortest path. There might be a path that is composed of more edges, but when those edges are shorter, the total path is as well. The path that needs to be found now needs to be calculated by adding all weights. A frequently used algorithm to handle weighted graphs is Dijkstra's algorithm (Dijkstra, 1959). Dijkstra's algorithm adds nodes with lowest costs first to its tree and explores along those nodes, updating earlier nodes if a shorter path

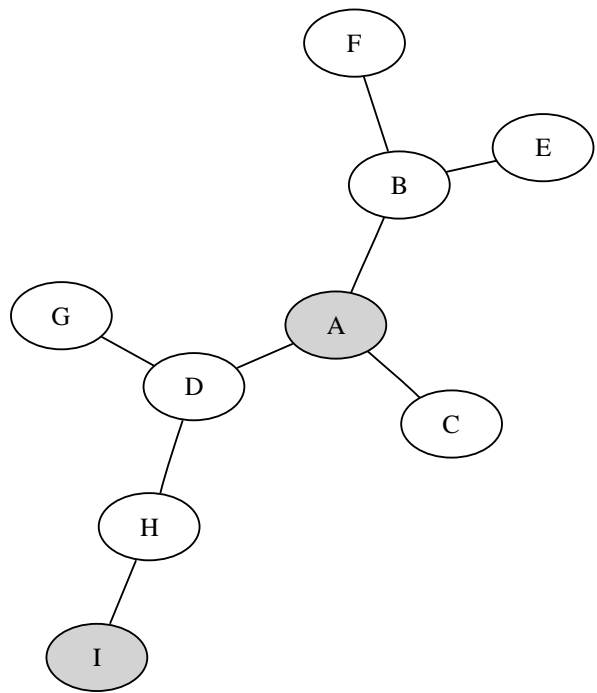


Figure 4.3: DFS and BFS traversing a network, with sourcenode *A* and targetnode *I*.

Step	DFS	BFS	Bidirectional
1	A	A	A
2	B	B	I
3	E	C	B
4	F	D	C
5	C	E	D
6	D	F	H
7	G	G	
8	H	H	
9	I	I	

is found. When a binary heap is used to store the tree in, time and space complexity for Dijkstra is $\mathcal{O}(|E| + |V| \log |V|)$.

A modification of Dijkstra’s algorithm wich generally reduces time and space complexity is A^* . This algorithm uses a heuristic to predict the direction it searches in (Stout, 1996). A^* is generally used in automobile- and web-based systems for computing driving directions and in pathfinding for non-player characters in videogames. It is also used in parsing, string matching, and structured prediction in computational linguistics.

A New Measure: Free Link Pathfinding

In this chapter I describe a new measure of calculating semantic relatedness by finding the shortest path between two concepts in an associative network. The associative networks that are explored are the link structure extracted out of Wikipedia and the predicates from the ConceptNet 3 database.

5.1 Creating the network

The Wikipedia dump contains over 2 million articles and 55 million links. This means the resulting graph will have over 2 million nodes and 55 million edges. In ConceptNet there are over 18 thousand concepts (nodes) that can be used (meaning they are not singletons) and over 254 thousand assertions (edges) that have a score higher than zero. Both networks are sparse: they do not nearly have as much connections as they could possibly have (namely, to all other nodes). To search these networks efficiently, they need to be indexed well and the algorithms need to be efficient.

5.1.1 Downloading the Wikipedia dump

If Wikipedia is regarded as a source to calculate semantic relatedness, it needs to be transformed into a semantic network. In this network the articles will become the nodes, and the hyperlinks between the articles can be considered as an establishment of a certain relation between the article it links from and the article it links to. The nature of the relation is not defined: in that regard it is similar to ConceptNet's *ConceptuallyRelatedTo* relation. The hyperlinks will be transformed into the edges of the network, connecting the articles with each other. Hyperlinks linking to pages outside Wikipedia are discarded as well.

To form the semantic network, the nodes and edges need to be extracted from Wikipedia. To do

this, first a dump of all the articles in Wikipedia needs to be downloaded. Regularly, Wikipedia dumps are made available for download in XML and SQL format on the Wikimedia download pages¹. For this thesis, the English Wikipedia dump dated 12 march 2008 containing all articles was downloaded². This zipped dump is 3.5 Gb big, and it unzips to one large XML file of approximately 15 Gb, containing all the Wikipedia articles. An example snippet of Wikipedia code taken from the first paragraph of the article *Baker* is below:

```
:''This article refers to the cooking profession. For other uses, see
[[Baker (disambiguation)]]''
[[Image:USS John C. Stennis baker.jpg|200px|thumb|right|A baker prepares
fresh rolls]]
```

```
A '''baker''' is someone who primarily [[bake]]s and sells [[bread]].
[[Cake]]s and similar foods may also be produced, as the traditional
boundaries between what is produced by a baker as opposed to a
[[pastry chef]] have blurred in recent decades. The place where a baker
works is called a '''bakehouse''', '''bakeshop''', or '''[[bakery]]'''.
```

5.1.2 Extracting the link structure

To form the network, only the article-names and hyperlinks between the articles need to be extracted. This is essentially the link structure. The remaining free text can be discarded entirely. The only information that is stored is the type of link and the paragraph it was found in. A link can be a normal link, appearing in the free text, or it can be either a category, redirect or recommendation as described in Chapter 2. In the extraction phase, one text-file is created containing all article names and internal hyperlinks along with information about those links in that article. This text-file is 1.8 Gb in size.

5.1.3 Indexing in- and outlinks

The extracted file containing the links is far too big to allow efficient searching. Therefore, it needs to be split into smaller files. Another problem that arises is that the file only contains the outgoing links from any given article. It is however necessary to also know the incoming links to any given article. To find these, an inverted index needs to be made of the entire extracted link structure file. This is done by reading the entire file into memory and then maintaining a hash of articles and links to that article. Every time a link to a certain article *X* is read from the file, the corresponding article the links originates from is added to the list of articles linking to *X*. If we look at the example

¹<http://download.wikimedia.org/>

²<http://download.wikimedia.org/enwiki/20080312/enwiki-20080312-pages-articles.xml.bz2>

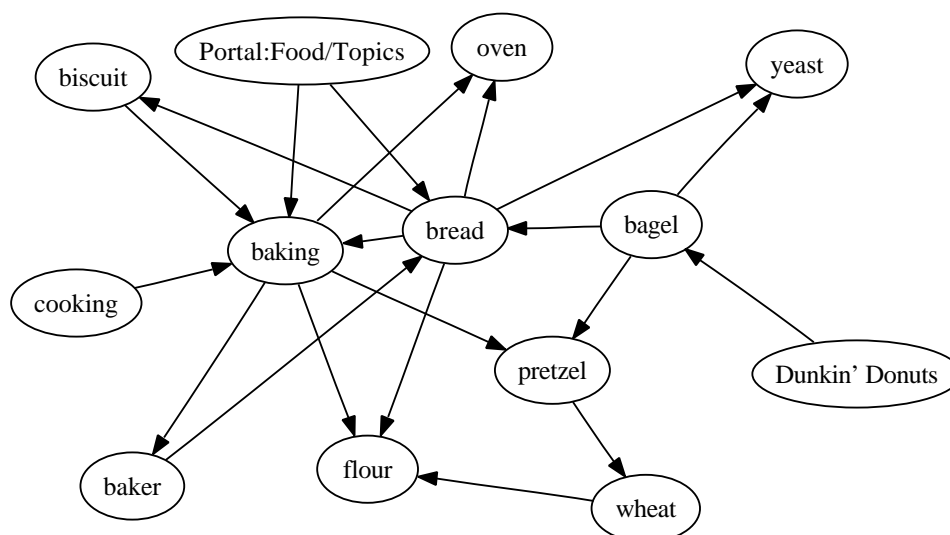


Figure 5.1: An example of some of the links found in various pages of Wikipedia.

in Figure 5.1, the outgoing links for *bread* (*oven*, *biscuit*, *yeast* and *flour*) can easily be extracted, but the incoming links need to be collected from the pages of *baker*, *bagel* and *Portal:Food/Topics*.

The incoming and outgoing links are stored in different directories. It would be optimal to store each concept in its own file. The downside to this approach is that it uses massive amounts of disk space. Therefore, I opted for the approach of storing the concepts in files based on the four first letters of the word. The file *Circ.txt* in the *index_to-directory* would for example hold among others the links leading to *Circle*, *Circus*, *Circulation* etcetera. Figure 5.2 offers an overview of the extraction process and structure of the index that is made.

5.1.4 Using ConceptNet 3

ConceptNet 3 can be freely downloaded as a PostgreSQL database, but also in Notation3 (N3) format. N3 is a compact and human readable non-XML serialization of Resource Description Framework models. I downloaded the N3 file³ from the ConceptNet website. Due to the database being in N3 format, the links can be extracted easily, along with user-assigned scores of the relationships. Because users have the option to input text in natural language, the data needs some processing in order to be of any value. An example of the N3 notation of ConceptNet 3 is below.

```

<http://conceptnet.media.mit.edu/assertion/1115669>
conceptnet:LeftConcept <http://conceptnet.media.mit.edu/concept/1000827>;
conceptnet:RelationType <http://conceptnet.media.mit.edu/reltype/UsedFor>;
conceptnet:RightConcept <http://conceptnet.media.mit.edu/concept/1002231>;
conceptnet:LeftText "a knife";
conceptnet:RightText "cutting";
conceptnet:FrameId <http://conceptnet.media.mit.edu/frame/1441>;

```

³http://conceptnet.media.mit.edu/conceptnet_en_20080605.n3.bz2

```

conceptnet:Language <http://conceptnet.media.mit.edu/language/en>;
conceptnet:Creator <http://conceptnet.media.mit.edu/user/10889>;
conceptnet:Score 50;
conceptnet:Sentence "a knife is used for cutting.".

```

5.1.5 Normalization

The assertion above defines the relation between *a knife* and *cutting*. It could just as easily have been a relation between *knives* and *to cut*. Although the form can be different, the concepts are the same. This requires the normalization of concepts: *knives*, *a knife*, *the knife* etcetera all need to be normalized to one concept. This normalization step takes place before the concepts can be indexed. Normalization requires two actions: first, non-content words need to be removed. These stopwords are words such as *a*, *the*, *many*, *be* and so on. Then, the words that remain need to be stemmed. This means that all suffixes are stripped so that only the stem remains. For this task, a Perl implementation of the Porter stemmer⁴ (Porter, 1997) is used. The stopwords-list that is used is from Snowball⁵. After normalization, the same indexing process can be used for ConceptNet as was used for the Wikipedia link structure. The extracted and normalized text file containing all concepts, scores and links is only 4 Mb large.



Figure 5.2: Overview of the extraction process

⁴<http://www.ldc.usb.gov/vdaniel/porter.pm>

⁵<http://snowball.tartarus.org/index.php>

5.2 Implementing the search-algorithm

Because of the richness and scale of the links in small-world networks, a depth-first search is not very likely to yield any results fast. The number of branches grow exponentially after each step, meaning that finding the right branch is nearly impossible using depth-first search. Breadth-first search is a better option, although in large networks such as Wikipedia performance can become a problem: the front of links that need to be followed can become very large. If Wikipedia is indeed a scale-free network this problem can be overcome. First of all, we know that in that case the maximum number of steps stays constant, no matter how large the network is. And secondly, because both in- and outgoing links were indexed, both forward and backward chaining can be used, meaning a bidirectional breadth-first search algorithm can be applied. This limits the time and space complexity considerably, as was shown in Chapter 4. For the implementation of the bi-directional breadth-first search algorithm, the `Algorithm::Sixdegrees` module⁶ is used. Figure 5.3 illustrates how the search process works.

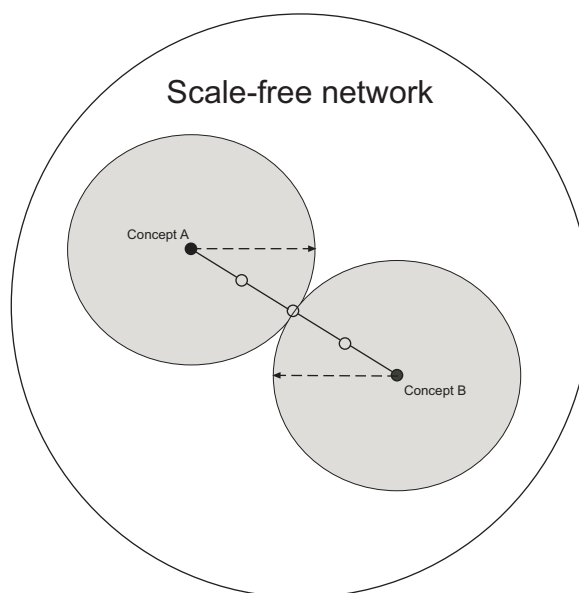


Figure 5.3: A schematic of bidirectional BFS in a scale-free network. The grey areas indicate the space that has been searched. The chain that connects both circles is a shortest path

5.2.1 Calculating relatedness

Semantic relatedness is calculated by using the total length of the shortest path that is found between two concepts c_1 and c_2 . Similar to the basic path-measure in WordNet, the number of nodes Np that are in the path are counted. Free link relatedness is then calculated as follows:

$$relatedness_{freelink}(c_1, c_2) = \max[\frac{1}{Np}]$$

When c_1 and c_2 are the same, only one node is in the path, and the relatedness will thus be $\frac{1}{1} = 1$.

⁶<http://search.cpan.org/~petek/Algorithm-SixDegrees-0.03/lib/Algorithm/SixDegrees.pm>

For every extra node in the path, the semantic relatedness decreases.

5.2.2 Directed search

Because links in Wikipedia are directed, it is logical to treat them as such. This means the forward chain only follows out-links and the backward chain only follows in-links. Once both chains are connected a path is found. Because all links are followed, the first connection that is found is always a shortest one. However, it is possible that the path leading from concept *A* to concept *B* is longer than the path from concept *B* to concept *A*. Therefore, both the path from *A* to *B* and the path from *B* to *A* needs to be found, and then the shortest path is selected.

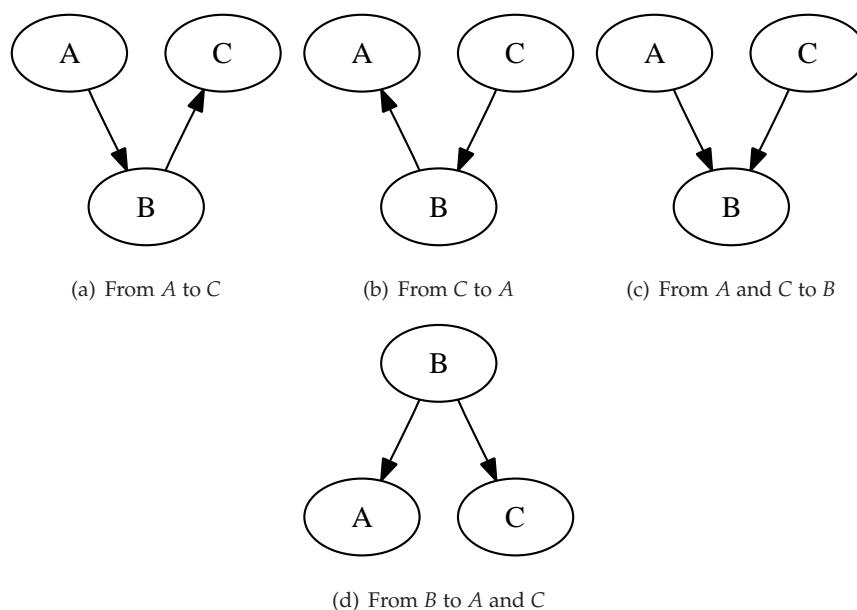


Figure 5.4: Four different ways of connection between node *A* and *C* through *B*. 5.4(a) and 5.4(b) are discovered when using directed search, 5.4(c) and 5.4(d) only when using undirected search

5.2.3 Undirected search

In ConceptNet, links do not have a direction. A predicate indicates a relationship between two concepts. This relation is thus true for both concepts. In the same fashion, although links in Wikipedia are directed, it might be beneficial to regard them as having no direction. If concept *A* links to concept *B* it could be argued not only is *A* related to *B*, apparently *B* does have some relation to *A* as well. Furthermore, if there is a concept *B* that links to both *A* and *C*, or *A* and *C* both link to *B*, this relationship could never be found by following directed links, but it could be found by not regarding links as having a direction (Figure 5.4). The page about *cars* might for example link to both *Ferrari* and *Lamborghini*, indicating some relation between the two. To implement undirected search, nodes are expanded by adding all in- and outlinks. This is done for both the forward and

the backward chain. Again, when both chains connect, a shortest path is found.

5.2.4 Weighted search

The scores that are assigned by users to assertions in ConceptNet can be used to apply scores to edges. This and the fact that the network is relatively small means weighted search can be applied to ConceptNet. An implementation of Dijkstra's algorithm is used for this task. *A** is not an option, because in a conceptual network it is very hard to apply a heuristic to determine the general search direction. For the Dijkstra implementation the `Boost::Graph` module⁷ is used. The semantic distance is calculated by assigning the inverse of the scores assigned to the predicates as costs to the edges and then finding the path with the lowest cost:

$$distance_{freelinkweighted}(c_1, c_2) = \min[\sum_{i=m}^n \frac{1}{s_i}]$$

To get the semantic relatedness, the inverse of this score is taken:

$$relatedness_{freelinkweighted}(c_1, c_2) = \frac{1}{distance_{freelinkweighted}(c_1, c_2)}$$

⁷<http://search.cpan.org/~dburdick/Boost-Graph-1.4/Graph.pm>

Experiments

Evaluating a system that calculates semantic relatedness is not an easy task. There is no universal truth that determines how related two concepts are. Judging semantic relatedness is typically a human task, because we do this automatically every day. The best way to evaluate such a system is therefore to compare it to how humans would do given the same task. This can be done by collecting human judgements for a representative sample of word pairs in an experimental setting. The average judgements can then be used for automatic evaluation. There are not many datasets that are based on semantic relatedness. Rather, most datasets focus on semantic similarity (i.e. how synonymous two words are). Examples of these are the Rubenstein and Goodenough (Rubenstein and Goodenough, 1965) and Miller and Charles (Miller and Charles, 1991) word pairs.

6.1 The Finkelstein WordSimilarity-353 test collection

The Finkelstein-353¹ test collection is a dataset that does contain semantic relatedness scores. In addition to that, it also is a very large dataset. While the previously mentioned Rubenstein & Goodenough dataset only contains 65 word pairs, the Finkelstein set contains 353 word pairs, among which are the 30 word pairs from the Miller & Charles dataset, but with newly assigned judgements.

The collection contains two sets of English word pairs along with human-assigned similarity judgements. The first set contains 153 word pairs along with their similarity scores assigned by 13 subjects. The second set contains 200 word pairs, with their similarity assessed by 16 subjects. All the subjects in both experiments were Israeli students who possessed near-native command of English. Their instructions were to estimate the relatedness of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words) (Finkelstein et al., 2002). The subjects were specifically instructed to take into account all possible relations, even antonymy. On average, judgements of individual subjects show a Spearman's rank order $\rho = 0.79$ with the whole group. Table 6.1 displays some examples of these word pairs. This dataset is used as a gold

¹<http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html>

standard to evaluate the different measures of calculating semantic relatedness.

mile	kilometer	8.66
computer	news	4.47
territory	surface	5.34
atmosphere	landscape	3.69
president	medal	3.00
war	troops	8.13
record	number	6.31
skin	eye	6.22
Japanese	American	6.50
theater	history	3.91
volunteer	motto	2.56
prejudice	recognition	3.00
decoration	valor	5.63
century	year	7.59
century	nation	3.16
delay	racism	1.19
delay	news	3.31
minister	party	6.63
peace	plan	4.75
minority	peace	3.69
attempt	peace	4.25
government	crisis	6.56
deployment	departure	4.25

Table 6.1: A collection of word pairs from the Finkelstein-353 dataset along with the averaged assigned scores.

6.2 Correlation

The data from the Finkelstein dataset and the scores assigned by the different metrics is compared using Spearman's ρ rank order coefficients ([Spearman, 1904](#)). This is a non-parametric method of calculating correlation, meaning it does not make any assumptions about the distribution of the data. Because output-scores from the pathfinding measures are logarithmic and the scores assigned by humans are continuous, it is best to calculate correlation on ranks instead of raw scores.

6.3 WordNet experimental setup

To use the WordNet measures, the `WordNet::Similarity` package² is implemented (Pedersen et al., 2004), using WordNet 3.0. This module is used to find the distances between the concepts in the Finkelstein dataset, resulting in a score for each word pair. For information content metrics, it uses the British National Corpus (World Edition), the Penn Treebank (version 2), the Brown Corpus, the complete works of Shakespeare, and SemCor.

6.4 Free link pathfinding experimental setup

6.4.1 Scale-freeness

First of all, the distribution of connections is measured for both Wikipedia as ConceptNet to investigate the scale-freeness of both networks. Then, the developed free link pathfinding software is used on both the Wikipedia as the ConceptNet data to find semantic relatedness.

6.4.2 Pathfinding

The word pairs from the Finkelstein dataset are fed to the free link pathfinding system, resulting in scores for each pair. For Wikipedia both directed and undirected search are tested, for Conceptnet undirected search, as there are no obvious directions in the ConceptNet database, and weighted search using the Dijkstra algorithm. The scores that are used are the inverse scores as assigned by users of Open Mind Commons, because edges corresponding with highly rated assertions should have short lengths (the distance between the two concepts is apparently low).

6.4.3 Modifying the network

On top of that, reranking is done for Wikipedia, counting redirects as having only half the lengths of normal links. It is also tested how well the metric performs if only the links in the first paragraph of Wikipedia articles are used. The distribution of connections is measured for both Wikipedia as ConceptNet to investigate the scale-freeness of both networks.

²<http://search.cpan.org/dist/WordNet-Similarity/>

Results

7.1 WordNet based measures

The correlation of the different measures applied to WordNet 3.0 on the full Finkelstein dataset with human judgement can be observed in Table 7.1. Spearman rank-order correlation coefficient is used to compare the computed relatedness scores with the scores assigned by the human test subjects. Only one concept was not found in WordNet: *Maradona* did not yield any results. In Figure 7.1 are scatterplots which show the correlation between the ranks of human judgements and WordNet measures.

graph based	path	0.29
	lch	0.30
	wup	0.33
information content based	res	0.33
	lin	0.20
	jc	0.18
text overlap based	lesk	0.41
	vector	0.45

Table 7.1: Spearman's ρ rank order coefficients of WordNet measures with human judgements

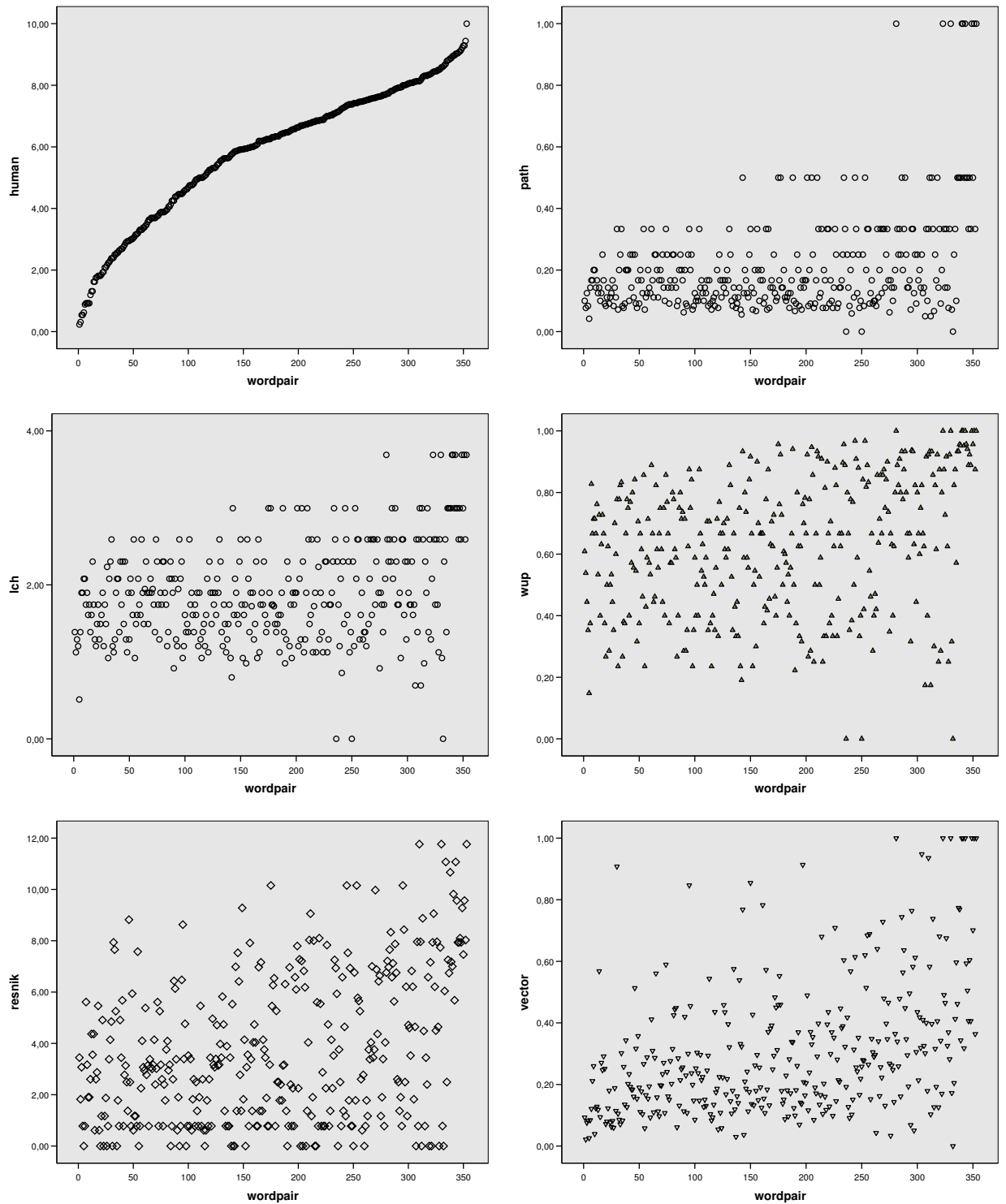


Figure 7.1: Visualisation of different WordNet metrics (shown on the Y-axis) ordered by rank assigned by human subjects

7.2 Free link pathfinding

7.2.1 Link distributions

The distributions of both in- and outdegree for all pages in Wikipedia is shown in Figure 7.2. It shows the relation between the number of incoming or outgoing links for a page and the total number of times a page with that amount of incoming and outgoing links is found. The scale of both axes is logarithmic. The total connectedness of both Wikipedia (both in- and outdegree for a page) and ConceptNet 3 is displayed in Figure 7.3.

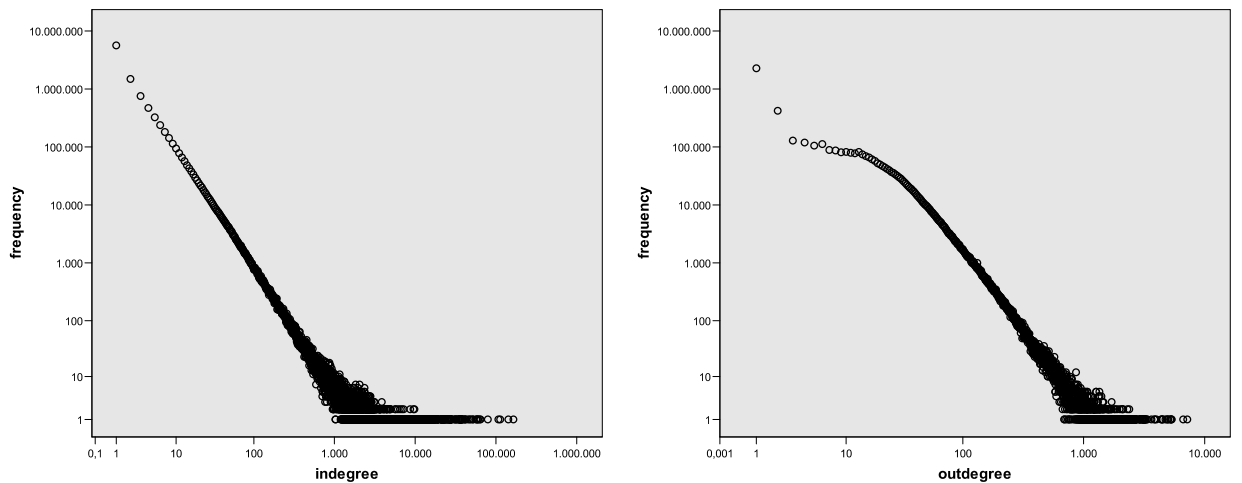


Figure 7.2: Logarithmic distribution of in- and outdegree for the pages found in Wikipedia

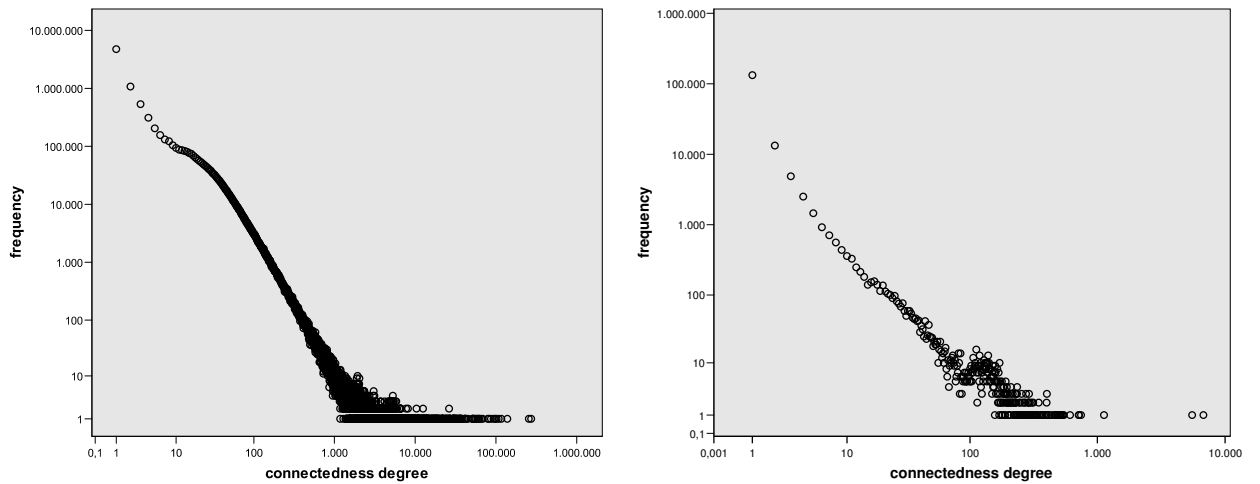


Figure 7.3: Logarithmic distribution of connectedness of both Wikipedia (left) and ConceptNet (right)

7.2.2 Relatedness

The basic directed free link pathfinding measure applied on Wikipedia shows a Spearman's ρ of **0.45** and the undirected measure shows a ρ of **0.56** with human judgements on the full Finkelstein dataset (Table 7.2). Only one wordpair could not be found: the word *defeating* was not found. Reranking the undirected results by giving redirect links a weight of 0.5 and all other links a weight of 1 resulted in $\rho = 0.56$. When using only the links found in the first paragraph of an article the correlation is $\rho = 0.51$ when using undirected search. Figure 7.4 displays scatterplots which show the correlation between the rank of human judgements and those of the pathfinding measures.

FLP Wikipedia directed	0.45
FLP Wikipedia undirected	0.56
FLP Wikipedia first paragraph undirected	0.51
FLP Wikipedia undirected reranked	0.56
FLP ConceptNet	0.35
FLP ConceptNet weighted	0.21
FLP ConceptNet non-missing	0.47

Table 7.2: Spearman's ρ rank order coefficients of different free link path (FLP) variants with human judgements.

WordNet measures	0.18 - 0.45
WikiRelate!	0.19 - 0.48
ESA Wikipedia	0.75
FLP Wikipedia	0.56
FLP ConceptNet	0.35

Table 7.3: Spearman's ρ rank order coefficients of different measures with human judgements.

In ConceptNet, 58 of the wordpairs could not be found, significantly impairing the score. The free link pathfinding measure applied to ConceptNet 3 yields a ρ of **0.35**. When we look only at those wordpairs that were found, the measure applied to ConceptNet performs slightly better than the directed Wikipedia measure, but worse than the undirected Wikipedia measure: it shows a ρ of **0.47**. The Dijkstra algorithm using the assigned scores by humans as costs gave a result of $\rho = 0.21$. In Table 7.4 some examples are shown of the paths found. The size of the ConceptNet indexes is very small compared to the indexes generated from Wikipedia: in total only 69 Mb versus the 3.4 Gb of Wikipedia. This means the runtime of the search algorithm is significantly shorter on the ConceptNet data than on the Wikipedia data. A matter of seconds versus minutes. A comparison between free link pathfinding and other measures is presented in Table 7.3. Results from ESA and WikiRelate! were taken from (Gabrilovich and Markovitch, 2007).

7.2.3 Path lengths

For all wordpairs found, the number of nodes visited was never more than 6, meaning only 5 hops were needed at most to connect any wordpair. On average, undirected search in Wikipedia needed 2.3 hops to connect the two words, undirected search in Wikipedia using only the first paragraph required 2.5 hops and the measure applied to ConceptNet connected two concepts in 2.4 hops. Undirected Wikipedia search needed significantly more hops: in 3.4 hops it was able to connect two concepts.

Wikipedia undirected
Racism <-> United States <-> Broadcast delay <-> Delay Government <-> Military <-> Crisis Cucumber <-> Agriculture <-> Potato Doctor <-> Doctors (BBC soap opera) <-> Nurse Smart <-> Genius <-> Stupidity <-> Stupid
Wikipedia directed
Racism -> European Union -> 1950s -> Delay (audio effect) -> Delay Crisis -> List of psychology topics -> Heuristic -> Social contract -> Government Cucumber -> Fruit -> Potato Doctor -> Mental health professional -> Psychiatric and mental health nursing -> Nurse Smart -> Catholic Bishops'... -> Tagalog... -> Grammatical number -> Noun -> Stupid
ConceptNet
Racism <-> Pain <-> Person <-> Delay Government <-> Person <-> Pray <-> Crisis Cucumber <-> Farmer's market <-> Potato Doctor <-> Nurse Smart <-> People <-> Stupid

Table 7.4: Some examples of paths found between concepts in Wikipedia and ConceptNet

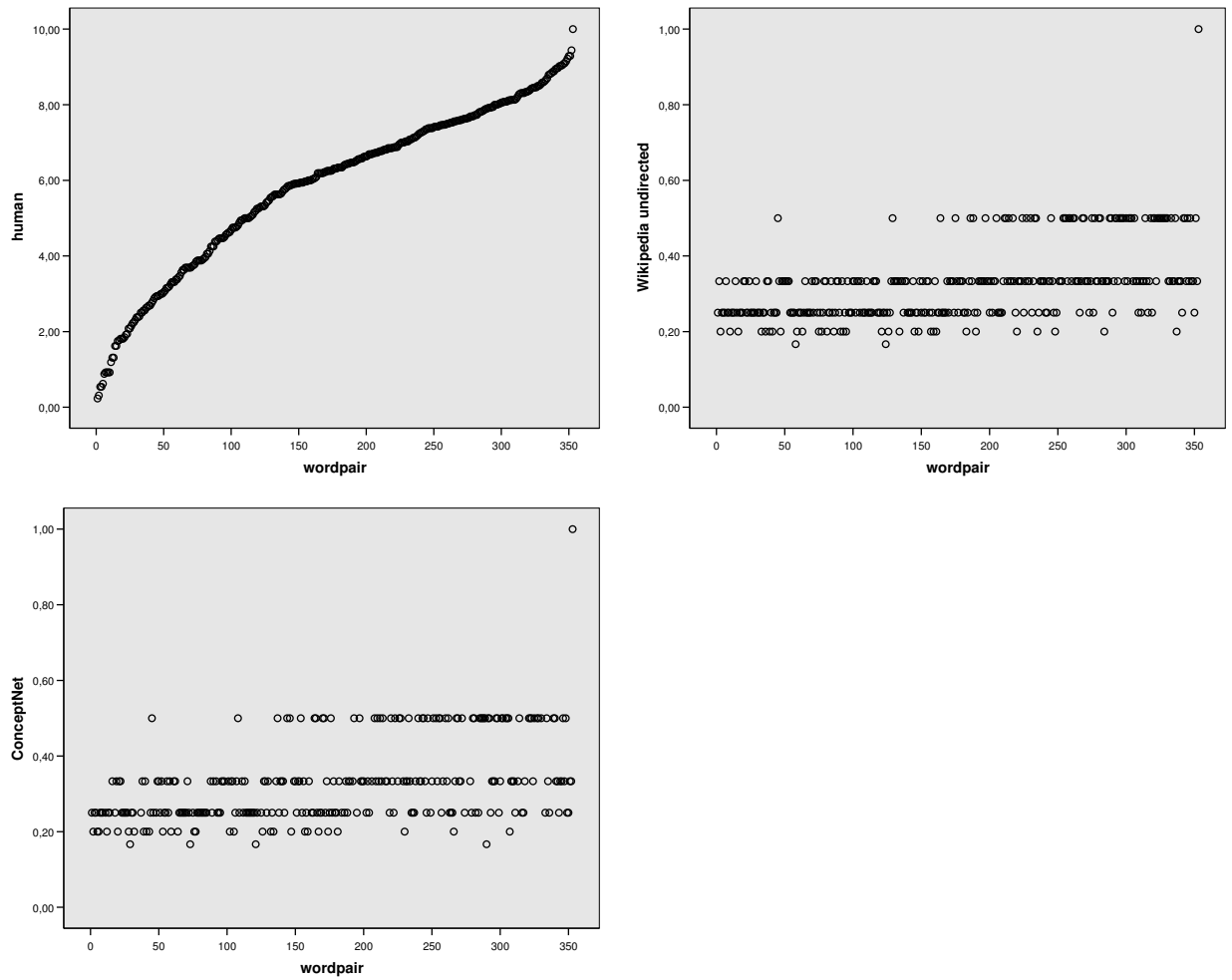


Figure 7.4: Visualisation of the free link pathfinding metric used on different resources (shown on the Y-axis) ordered by rank assigned by human subjects

Discussion

In this Chapter the research questions posed in Chapter 1 are addressed. The results obtained in Chapter 7 are discussed to draw a conclusion.

8.1 Network type

Both ConceptNet and Wikipedia networks are constructed using free link structure. Link distributions in both networks seem to follow a powerlaw. In particular the indegree of Wikipedia articles shows a very clear power law distribution, only fanning out near the bottom of the graph where frequencies are very low. This is because frequencies lower than 1 cannot exist for numbers of links leading to existing pages, and distributions of frequencies of these highly connected pages are somewhat random. The outdegree shows a sort of concave gap at the top of the graph, roughly between 1 and 20 outlinks. This means in Wikipedia articles with 1 to 20 outgoing links are less frequent than would be expected from a scale-free network. The constraints Wikipedia endorses on its users might have something to do with this: articles with few links are more likely to be flagged for expansion, deletion or merging with existing articles. Apparently a typical Wikipedia article links to 20 or more other articles. Apart from this, the distribution is very similar to the in-rank distribution, albeit on a somewhat smaller scale: articles can have only that many outgoing links before they become too large to be readable and editable.

Overall, both Wikipedia and ConceptNet seem to show a distribution that roughly follows a power law distribution. ConceptNet's distribution seems a little more random, probably because of the significantly smaller scale of the network. The average path-lengths on the finkelstein dataset seem to confirm this notion. On networks of three different scales (namely the entire Wikipedia, only the first paragraph of Wikipedia and ConceptNet), the average number of hops needed to get from one word to the other does not differ very much: respectively 2.3, 2.5 and 2.4.

8.2 ConceptNet versus Wikipedia

With a correlation of 0.56 for undirected search versus 0.45 for directed, undirected is the best way to go when computing semantic relatedness using the Wikipedia network, despite the fact that links in Wikipedia do have an explicit direction. This can probably be accounted to the more complex type of relations that can be covered by undirected search as described in Chapter 5. Wikipedia undirected search also easily outperforms ConceptNet, which only scores a correlation of 0.35. This lower score can mainly be attributed by the lack of coverage of ConceptNet. In Wikipedia, only 1 wordpair could not be found, while in ConceptNet 58 could not be found. Even when only non-missing wordpairs are considered, ConceptNet performs worse than Wikipedia with a correlation of 0.47 with human judgements. It is better than Wikipedia directed search, so it definitely shows potential, taking into account that the ConceptNet indexes are roughly 50 times smaller than the ones created from Wikipedia.

Using only the first paragraph of articles in Wikipedia, which generally contains the definition of a word, still generates satisfying results ($\rho = 0.51$), but is still lower than when the entire Wikipedia is used. Apparently there are more useful links in the body of the article than there are ‘garbage’ links that mess up the calculation of semantic relatedness. Still, the first paragraph shows to be a good representation of an article. Reranking results using the notion that redirect links cost half as much as normal links does not change the score significantly. Apparently it does as much good as it does damage. The idea is not wrong: redirects show a very strong relationship between two pages. The problem here is the implementation: because of the richness of the network, the algorithm arbitrarily can take a number of routes between concepts which are all equal in length. If redirects are to be considered as having a lower cost than normal links, then it needs to be made sure that the algorithm always prefers to take redirect routes rather than normal links. For this, weighted search needs to be used. In many experiments, weighted search did not provide better results for ConceptNet. Results using weights were actually worse ($\rho = 0.21$). However, weighted search was not the focus of this study. The scores can probably be used in another way to augment the pathfinding scores, maybe by combining the number of hops and the scores in another way than multiplying them.

8.3 Free link pathfinding versus other methods

The free link pathfinding method using Wikipedia as introduced in this thesis outperforms any other existing pathfinding method for calculating semantic relatedness. It also outperforms any method that makes use of WordNet, even the ones that make use of textual content, such as extended gloss vectors. This cannot be explained by coverage: both in WordNet 3.0 and in the Wikipedia dump of march 2008 only one wordpair was not found. This proves that a bottom-up free link structure in conceptual networks is better for finding semantic relatedness than top-down hierarchical structures as used in ontologies, such as WordNet.

The performance of WikiRelate! is also lower, but it needs to be taken into account that the Wikipedia dump that was used in this thesis is a lot newer and bigger than the dump used for WikiRelate!. Still, when only similar pathfinding metrics are considered, the performance achieved by free link pathfinding scores significantly better. While ESA's performance ($\rho = 0.75$) is even higher, the comparison is not realistic. ESA is a fully integrated machine learning architecture that makes extensive use of Wikipedia's free text. The free link measure is a robust, simple and portable method that uses only link structure from any given conceptual network.

8.4 Future research

I recommend using the Wikipedia free link pathfinding algorithm for doing future research into semantic relatedness, as it has proved to be a valid method of calculating the relatedness of concepts and preferable above methods that use a hierarchical knowledge structure. Although a first version was described in this thesis, I believe the free link pathfinding algorithm can be thoroughly improved. This initial version was developed to be adapted easily to different conceptual networks, and to treat all links in that network as being equal. Although I experimented with weighted search, no satisfying results were yet obtained. It is however my belief that this approach should be pursued further, to find the proper way to obtain and use scores to weigh relations. This could result in obtaining more accurate and diverse results, as opposed to the discrete scores that are obtained now. On the Finkelstein-353 dataset, only six different scores were obtained for relatedness of all word pairs (see Figure 7.4, due to the limited amount of steps needed in a scale-free network).

It is also recommendable to evaluate the free link pathfinding measure by using it in other higher order natural language processing tasks, such as automatic summarization, information retrieval, word sense disambiguation and machine translation. It should be valuable to explore how much these tasks can be improved by using this method to calculate semantic relatedness.

References

- Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco.
- Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.
- Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-free brain functional networks. *Phys Rev Lett*, 94(1).
- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131.
- Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.
- Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.
- Havasi, C., Speer, R., and Alonso, J. (2007). Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria.
- Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X*, Taiwan.
- Kozima, H. and Furugori, T. (1993). Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 232–239, Morristown, NJ, USA. Association for Computational Linguistics.

- Leacock, C., Chodorow, M., and Miller, G. A. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165.
- Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Lenat, D. B. and Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA. ACM.
- Leuf, B. and Cunningham, W. (2001). *The Wiki way: quick collaboration on the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.
- Liu, H. and Singh, P. (2004). Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.
- Mandala, R., Tokunaga, T., Tanaka, H., Okumura, A., and Satoh, K. (1998). Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. In *Text REtrieval Conference*, pages 414–419.
- Milgram, S. (1967). The small-world problem. *Psychology Today*, (1):61–67.
- Miller, G. and Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.
- Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.
- Minsky, M. (1986). *The society of mind*. Simon & Schuster, Inc., New York, NY, USA.
- Moldovan, D. I. and Mihalcea, R. (2000). Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43.
- Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.
- Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.
- Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity - measuring the relatedness of concepts.
- Porter, M. F. (1997). An algorithm for suffix stripping. pages 313–316.

- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference for Artificial Intelligence (IJCAI-95)*, pages 448–453.
- Richardson, R. and Smeaton, A. F. (1995). Using WordNet in a knowledge-based approach to information retrieval. Technical Report CA-0395, Dublin, Ireland.
- Riehle, D. (2006). How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 3–8, New York, NY, USA. ACM.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.
- Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
- Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1223–1237, London, UK. Springer-Verlag.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 100(3-4):441–471.
- Spek, S. (2006). Wikipedia: organisation from a bottom-up approach. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, New York, NY, USA. ACM.
- Stout, B. (1996). Smart moves: Intelligent pathfinding. *Game Developer Magazine*.
- Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *21. AAAI / 18. IAAI 2006*. AAAI Press.
- Viégas, F. B., Wattenberg, M., and Mckee, M. (2007). The hidden order of wikipedia. pages 445–454.
- Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.
- Voss, J. (2005). Measuring wikipedia. In *Proceedings of ISSI 2005*, Stockholm, Sweden.
- Vossen, P., editor (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.
- Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA. ACM.
- Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico State University, Las Cruces, New Mexico.

Appendix A

Below are the results from Wikipedia and ConceptNet undirected search for all wordpairs in the Finkelstein WordSimilarity-353 set.

word pair	human	Wikipedia	ConceptNet
love-sex	6.77	.33	.50
tiger-cat	7.35	.33	.50
tiger-tiger	10.00	1.00	1.00
book-paper	7.46	.50	.50
computer-keyboard	7.62	.33	.50
computer-internet	7.58	.50	.50
plane-car	5.77	.25	.33
train-car	6.31	.33	.33
telephone-communication	7.50	.33	.50
television-radio	6.77	.50	.33
media-radio	7.42	.33	.50
drug-abuse	6.85	.33	.33
bread-butter	6.19	.50	.50
cucumber-potato	5.92	.33	.33
doctor-nurse	7.00	.33	.50
professor-doctor	6.62	.33	.33
student-professor	6.81	.33	.33
smart-student	4.62	.25	.33
smart-stupid	5.81	.25	.33
company-stock	7.08	.33	.33
stock-market	8.08	.50	.50
stock-phone	1.62	.33	.25
stock-CD	1.31	.25	.20
stock-jaguar	.92	.33	.25
stock-egg	1.81	.25	.25
fertility-egg	6.69	.25	.25
stock-live	3.73	.25	.33
stock-life	.92	.25	.25
book-library	7.46	.50	.50
bank-money	8.12	.50	.50
wood-forest	7.73	.50	.50
money-cash	9.15	.50	.50
professor-cucumber	.31	.33	.20
king-cabbage	.23	.25	.25
king-queen	8.58	.50	.50
king-rook	5.92	.25	.33
bishop-rabbi	6.69	.33	.33
Jerusalem-Israel	8.46	.50	.50
Jerusalem-Palestinian	7.65	.50	.
holy-sex	1.62	.25	.
fuck-sex	9.44	.33	.33
Maradona-football	8.62	.33	.
football-soccer	9.03	.33	.33
football-basketball	6.81	.50	.50

REFERENCES

football-tennis	6.63	.33	.33
tennis-racket	7.56	.33	.25
Arafat-peace	6.73	.25	.
Arafat-terror	7.65	.25	.
Arafat-Jackson	2.50	.20	.
law-lawyer	8.38	.50	.50
movie-star	7.38	.33	.25
movie-popcorn	6.19	.25	.50
movie-critic	6.73	.25	.50
movie-theater	7.92	.33	.50
physics-proton	8.12	.50	.20
physics-chemistry	7.35	.50	.33
space-chemistry	4.88	.33	.25
alcohol-chemistry	5.54	.50	.25
vodka-gin	8.46	.50	.
vodka-brandy	8.13	.33	.33
drink-car	3.04	.33	.33
drink-ear	1.31	.25	.25
drink-mouth	5.96	.33	.50
drink-eat	6.87	.20	.33
baby-mother	7.85	.33	.50
drink-mother	2.65	.33	.33
car-automobile	8.94	.50	.50
gem-jewel	8.96	.25	.33
journey-voyage	9.29	.25	.25
boy-lad	8.83	.20	.
coast-shore	9.10	.50	.33
asylum-madhouse	8.87	.33	.
magician-wizard	9.02	.50	.25
midday-noon	9.29	.50	.33
furnace-stove	8.79	.50	.33
food-fruit	7.52	.50	.50
bird-cock	7.10	.50	.33
bird-crane	7.38	.25	.50
tool-implement	6.46	.50	.25
brother-monk	6.27	.50	.25
crane-implement	2.69	.20	.20
lad-brother	4.46	.20	.
journey-car	5.85	.25	.25
monk-oracle	5.00	.33	.
cemetery-woodland	2.08	.33	.25
food-rooster	4.42	.25	.33
coast-hill	4.38	.33	.33
forest-graveyard	1.85	.33	.33
shore-woodland	3.08	.33	.25
monk-slave	.92	.25	.25
coast-forest	3.15	.33	.33
lad-wizard	.92	.20	.
chord-smile	.54	.20	.25
glass-magician	2.08	.25	.25
noon-string	.54	.25	.25
rooster-voyage	.62	.25	.20
money-dollar	8.42	.33	.50
money-cash	9.08	.50	.50
money-currency	9.04	.50	.33
money-wealth	8.27	.50	.33
money-property	7.57	.33	.33
money-possession	7.29	.25	.33
money-bank	8.50	.50	.50
money-deposit	7.73	.33	.25
money-withdrawal	6.88	.33	.33
money-laundering	5.65	.25	.50
money-operation	3.31	.25	.33

REFERENCES

tiger-jaguar	8.00	.50	.33
tiger-feline	8.00	.50	.33
tiger-carnivore	7.08	.50	.50
tiger-mammal	6.85	.50	.33
tiger-animal	7.00	.50	.50
tiger-organism	4.77	.33	.33
tiger-fauna	5.62	.33	.20
tiger-zoo	5.87	.33	.50
psychology-psychiatry	8.08	.50	.
psychology-anxiety	7.00	.33	.
psychology-fear	6.85	.33	.
psychology-depression	7.42	.33	.
psychology-clinic	6.58	.33	.
psychology-doctor	6.42	.33	.
psychology-Freud	8.21	.33	.
psychology-mind	7.69	.50	.
psychology-health	7.23	.33	.
psychology-science	6.71	.50	.
psychology-discipline	5.58	.33	.
psychology-cognition	7.48	.50	.
planet-star	8.45	.50	.33
planet-constellation	8.06	.33	.33
planet-moon	8.08	.50	.50
planet-sun	8.02	.50	.33
planet-galaxy	8.11	.33	.50
planet-space	7.92	.50	.50
planet-astronomer	7.94	.50	.25
precedent-example	5.85	.25	.
precedent-information	3.85	.25	.
precedent-cognition	2.81	.25	.
precedent-law	6.65	.50	.
precedent-collection	2.50	.25	.
precedent-group	1.77	.25	.
precedent-antecedent	6.04	.20	.
cup-coffee	6.58	.50	.50
cup-tableware	6.85	.25	.25
cup-article	2.40	.25	.20
cup-artifact	2.92	.25	.25
cup-object	3.69	.25	.25
cup-entity	2.15	.25	.25
cup-drink	7.25	.33	.50
cup-food	5.00	.25	.33
cup-substance	1.92	.25	.33
cup-liquid	5.90	.25	.50
jaguar-cat	7.42	.33	.50
jaguar-car	7.27	.25	.33
energy-secretary	1.81	.25	.33
secretary-senate	5.06	.33	.25
energy-laboratory	5.09	.33	.25
computer-laboratory	6.78	.50	.50
weapon-secret	6.06	.25	.25
FBI-fingerprint	6.94	.50	.
FBI-investigation	8.31	.33	.
investigation-effort	4.59	.25	.33
Mars-water	2.94	.50	.50
Mars-scientist	5.63	.33	.25
news-report	8.16	.25	.25
canyon-landscape	7.53	.33	.25
image-surface	4.56	.33	.33
discovery-space	6.34	.33	.25
water-seepage	6.56	.33	.
sign-recess	2.38	.25	.25
Wednesday-news	2.22	.25	.25

REFERENCES

mile-kilometer	8.66	.50	.25
computer-news	4.47	.33	.33
territory-surface	5.34	.25	.25
atmosphere-landscape	3.69	.25	.25
president-medal	3.00	.33	.25
war-troops	8.13	.25	.33
record-number	6.31	.25	.25
skin-eye	6.22	.33	.25
Japanese-American	6.50	.33	.33
theater-history	3.91	.33	.25
volunteer-motto	2.56	.25	.
prejudice-recognition	3.00	.33	.33
decoration-valor	5.63	.20	.20
century-year	7.59	.50	.50
century-nation	3.16	.33	.20
delay-racism	1.19	.25	.25
delay-news	3.31	.25	.25
minister-party	6.63	.25	.33
peace-plan	4.75	.33	.33
minority-peace	3.69	.25	.25
attempt-peace	4.25	.25	.25
government-crisis	6.56	.33	.25
deployment-departure	4.25	.20	.
deployment-withdrawal	5.88	.20	.
energy-crisis	5.94	.33	.25
announcement-news	7.56	.25	.20
announcement-effort	2.75	.20	.20
stroke-hospital	7.03	.33	.33
disability-death	5.47	.33	.33
victim-emergency	6.47	.25	.33
treatment-recovery	7.91	.25	.17
journal-association	4.97	.25	.25
doctor-personnel	5.00	.25	.25
doctor-liability	5.19	.25	.25
liability-insurance	7.03	.50	.20
school-center	3.44	.25	.33
reason-hypertension	2.31	.33	.17
reason-criterion	5.91	.25	.20
hundred-percent	7.38	.20	.
Harvard-Yale	8.13	.33	.33
hospital-infrastructure	4.63	.25	.
death-row	5.25	.25	.25
death-inmate	5.03	.25	.25
lawyer-evidence	6.69	.33	.25
life-death	7.88	.50	.50
life-term	4.50	.20	.25
word-similarity	4.75	.25	.20
board-recommendation	4.47	.20	.25
governor-interview	3.25	.25	.25
OPEC-country	5.63	.33	.
peace-atmosphere	3.69	.25	.25
peace-insurance	2.94	.33	.25
territory-kilometer	5.28	.20	.17
travel-activity	5.00	.25	.33
competition-price	6.44	.50	.25
consumer-confidence	4.13	.33	.25
consumer-energy	4.75	.33	.33
problem-airport	2.38	.25	.25
car-flight	4.94	.25	.33
credit-card	8.06	.33	.50
credit-information	5.31	.25	.25
hotel-reservation	8.03	.33	.25
grocery-money	5.94	.25	.33

REFERENCES

registration-arrangemen	6.00	.20	.20
arrangement-accommodati	5.41	.20	.20
month-hotel	1.81	.33	.20
type-kind	8.97	.50	.33
arrival-hotel	6.00	.25	.25
bed-closet	6.72	.25	.33
closet-clothes	8.00	.50	.50
situation-conclusion	4.81	.25	.20
situation-isolation	3.88	.20	.25
impartiality-interest	5.16	.33	.25
direction-combination	2.25	.25	.20
street-place	6.44	.33	.33
street-avenue	8.88	.33	.50
street-block	6.88	.33	.25
street-children	4.94	.25	.50
listing-proximity	2.56	.20	.
listing-category	6.38	.20	.
cell-phone	7.81	.20	.25
production-hike	1.75	.20	.33
benchmark-index	4.25	.25	.
media-trading	3.88	.25	.20
media-gain	2.88	.25	.20
dividend-payment	7.63	.25	.
dividend-calculation	6.48	.25	.
calculation-computation	8.44	.50	.50
currency-market	7.50	.50	.33
OPEC-oil	8.59	.33	.
oil-stock	6.34	.33	.25
announcement-production	3.38	.20	.20
announcement-warning	6.00	.20	.20
profit-warning	3.88	.20	.20
profit-loss	7.63	.33	.25
dollar-yen	7.78	.33	.
dollar-buck	9.22	.50	.25
dollar-profit	7.38	.25	.25
dollar-loss	6.09	.25	.25
computer-software	8.50	.50	.50
network-hardware	8.31	.25	.25
phone-equipment	7.13	.20	.25
equipment-maker	5.91	.20	.
luxury-car	6.47	.20	.33
five-month	3.38	.25	.
report-gain	3.63	.25	.20
investor-earning	7.13	.25	.25
liquid-water	7.89	.50	.50
baseball-season	5.97	.33	.25
game-victory	7.03	.25	.33
game-team	7.69	.33	.33
marathon-sprint	7.47	.33	.25
game-series	6.19	.25	.25
game-defeat	6.97	.25	.33
seven-series	3.56	.20	.
seafood-sea	7.47	.50	.33
seafood-food	8.34	.50	.33
seafood-lobster	8.70	.50	.50
lobster-food	7.81	.33	.50
lobster-wine	5.70	.33	.25
food-preparation	6.22	.25	.50
video-archive	6.34	.25	.20
start-year	4.06	.25	.25
start-match	4.47	.25	.25
game-round	5.97	.25	.33
boxing-round	7.61	.33	.33

REFERENCES

championship-tournament	8.36	.50	.
fighting-defeating	7.41	.	.33
line-insurance	2.69	.25	.33
day-summer	3.94	.33	.25
summer-drought	7.16	.33	.25
summer-nature	5.63	.33	.33
day-dawn	7.53	.33	.25
nature-environment	8.31	.33	.25
environment-ecology	8.81	.50	.25
nature-man	6.25	.33	.50
man-woman	8.30	.50	.50
man-governor	5.25	.25	.25
murder-manslaughter	8.53	.50	.
soap-opera	7.94	.33	.33
opera-performance	6.88	.50	.50
life-lesson	5.94	.25	.33
focus-life	4.06	.25	.25
production-crew	6.25	.33	.25
television-film	7.72	.50	.50
lover-quarrel	6.19	.25	.20
viewer-serial	2.97	.20	.
possibility-girl	1.94	.25	.25
population-development	3.75	.33	.25
morality-importance	3.31	.17	.33
morality-marriage	3.69	.33	.25
Mexico-Brazil	7.44	.50	.33
gender-equality	6.41	.25	.25
change-attitude	5.44	.25	.33
family-planning	6.25	.33	.33
opera-industry	2.63	.33	.25
sugar-approach	.88	.25	.20
practice-institution	3.19	.25	.25
ministry-culture	4.69	.33	.
problem-challenge	6.75	.25	.33
size-prominence	5.31	.25	.
country-citizen	7.31	.33	.50
planet-people	5.75	.33	.33
development-issue	3.97	.20	.25
experience-music	3.47	.25	.33
music-project	3.63	.33	.25
glass-metal	5.56	.33	.33
aluminum-metal	7.83	.33	.50
chance-credibility	3.88	.25	.25
exhibit-memorabilia	5.31	.17	.
concert-virtuoso	6.81	.33	.
rock-jazz	7.59	.33	.33
museum-theater	7.19	.33	.33
observation-architectur	4.38	.33	.25
space-world	6.53	.33	.50
preservation-world	6.19	.25	.25
admission-ticket	7.69	.33	.25
shower-thunderstorm	6.31	.33	.33
shower-flood	6.03	.33	.33
weather-forecast	8.34	.25	.
disaster-area	6.25	.25	.20
governor-office	6.34	.25	.33
architecture-century	3.78	.33	.17

Appendix B

A collection of scripts developed for this thesis can be downloaded from:

<http://stuwww.uvt.nl/people/wubben/files/thesis.tar.gz>