Extracting Semantic Information from Wikipedia Using Human Computation and Dimensionality Reduction

Robert West

Master of Science

School of Computer Science

McGill University

Montreal, Quebec

2010-02-15

A thesis submitted to McGill University
in partial fulfillment of the requirements
of the degree of Master of Science

DEDICATION

Für Opa.

To my beloved late grandfather Heinrich Fleischer,

who loved playing with words and, a typesetter by trade,

would approve of my using Fraktur letters for variable names.

And to his daughter. And her husband. And their son. And his sister.

ACKNOWLEDGEMENTS

This work would not have been possible without the support and supervision of Joelle Pineau. She gave me the academic freedom that makes research fun, championed the scientific rigor that makes research good, and provided the financial support that makes research possible. I cannot envision a better supervisor.

I am equally obliged to Doina Precup, for always encouraging 'cool ideas' and advising me in much of the work presented in this thesis. I am thankful to both Joelle and Doina for fostering the friendly lab atmosphere that was a key component in making my time in Montreal as enjoyable as it has been, and to all members of the McGill Reasoning and Learning Lab: to \cos\min Păduraru for sharing my love of puns; to Fabian Kaelin for introducing me to Hoài Hương's spicy beef soup; to Amin Atrash for always reminding me that sampling is the solution to everything; to Rob Kaplow for supplying the lab with chocolate and software support; to Julien Villemure for debugging my machine code; to Mahdi Milani Fard, Ghiță Comanici, Jordan Frank, Monica Dinculescu, and Pablo Castro for coalescing in mission COMP-690; to Keith Bush and Arthur Guez for helping to rid the world of a dire disease; to Bert Vincent and Phil Bachman for their balanced *Weltanschauungen;* to Stéphane Ross for his jolly laughter; and to Mathieu Petitpas for taking big steps in advancing the Hackers hockey team.

Grandescunt aucta labore, but there is more than just work. I am grateful to the friends without whom my life in Montreal would have been so different. Foremost, I have to thank the dear people with whom I have been living in our Castle: Vince for showing me that fixing things can be fun and that food need not be bought; Gabie for her smiles; Laura for unforgettable minutes of unreined laughter;

Alissa—dik dik!—for her scones; Harlan for co-founding Eigenbräu and pinpointing the importance of the questionability–longevity trade-off; Maxi for inventing the avocado bagel sandwich; Matt for irie vibrations and bass; and Mark—big up! I also say *merci* to Imad for cultivating the Lebanese in me; to Patrick for having stopped offering me shots; to Yan, who, without even knowing me, hosted me when I first arrived in Montreal and has since become a good friend and brewing partner; to Peter for epic exertions by bike, ski, and foot; and to Katie for bringing ginger to my life.

Neither do I want to miss thanking all the anonymous Wikispeedia players, without whose contribution this research would have been impossible and pointless.

I would also like to usurp this opportunity to thank some people for things that have nothing to do with this work: My Ingolstadt homeboys have been a shaping force since our childhood, and I say *danke* to Aussi for giving better Rambo impressions than Sly Stallone; to Gbrl for going with me through the ordeal of my first dancing lessons; to Klausi for his adventurer's spirit; to Kulzi for the nice weather every April 2; to Schieder for sharing my passions for jazz and Bud Spencer movies; to Andy for memorable carnivals in Starnberg; and to Katha, not a homeboy but still one of us, for reading *Driss Ben Mohammed* out to me. Further south in Bavaria, I thank Küken for showing me Squidrullu, and Colosso for unrivaled inspiration and for Sankt-Peterburg. I am also obliged to Megan for standing by me when we were trapped on the Serbian highway.

Last and most I must thank my family. These few lines cannot capture how grateful I am to them. I say thank you to my Mama and Papa for loving me and caringly supporting me in all I do, and for letting me scribble around in our encyclopedia as a toddler. *Dühj* to my dear sister Tina, for Hasi and Heinerle and all they stand for. Finally, I thank my grandparents, Fleischer-Oma, West-Oma, West-Opa, and Fleischer-Opa, for always being around and always wanting me to stay.

ABSTRACT

Semantic background knowledge is crucial for many intelligent applications. A classical way to represent such knowledge is through semantic networks. Wikipedia's hyperlink graph can be considered a primitive semantic network, since the links it contains usually correspond to semantic relationships between the articles they connect. However, Wikipedia is rather noisy in this function. We propose Wikispeedia, an online human-computation game that can effectively filter this noise, furnishing data that can be leveraged to define a robust measure of semantic relatedness between concepts. While the resulting measure is very precise, it has the limitation of being sparse, i.e., undefined for many pairs of concepts. Therefore, we develop algorithms based on principal component analysis to increase coverage to the set of all pairs of Wikipedia concepts. These methods can also be generalized to other sparse measures of semantic relatedness, which we demonstrate by applying our approach to the Wikipedia adjacency matrix. Building on the same techniques, we finally propose an algorithm for finding missing hyperlinks in Wikipedia, which results in increased human usability.

ABRÉGÉ

Des connaissances d'arrière-plan sémantiques sont essentielles pour de nombreuses applications intelligentes. Les réseaux sémantiques constituent une façon classique de représenter de telles connaissances. On peut comprendre le graphe défini par les hyperliens de Wikipédia comme un réseau sémantique primitif, car les liens qu'il contient correspondent habituellement à des relations sémantiques entre les articles qu'ils joignent. Cependant, si on considère Wikipédia comme un réseau sémantique, le niveau de bruit est relativement élevé. Nous proposons Wikispeedia, un jeu de calcul humain en ligne qui peut effectivement filtrer ce bruit, en fournissant des données que nous utilisons pour définir une mesure de proximité sémantique entre les concepts. Bien que la mesure qui s'ensuit soit très précise, elle est creuse, c'est-à-dire indéfinie sur de nombreuses paires de concepts. Pour couvrir l'ensemble de toutes les paires de concepts que contient Wikipédia, nous développons des algorithmes basés sur l'Analyse en composantes principales. Ces méthodes peuvent être généralisées aux autres mesures de proximité sémantique creuses, ce que nous démontrons en appliquant notre approche à la matrice d'adjacence de Wikipédia. Enfin, nous utilisons les mêmes techniques en proposant un algorithme qui est capable de trouver les liens manquants dans Wikipédia, donnant lieu à un système de meilleure convivialité.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

'Time flies like an arrow.'—

Phrases like this are abundant whenever humans communicate. In fact, they come so naturally to us that we are most often not even aware of their complexity: to understand this saying, we must know that arrows can fly; that they tend to be very fast once they do so; that humans perceive the passing of time subjectively, sometimes as slower, sometimes as faster; and that humans often metaphorically map the passing of time to movement through space.

1.1 Common-Sense Knowledge

All humans—at least within a given culture—share vast amounts of such *common-sense knowledge,* to a much higher degree than they share, for instance, professional knowledge. Without it, we could not interact with one another.

As part of our common-sense reasoning, we continuously assess how closely related concepts are semantically. The capability to do so comes in handy, for instance, when spoken words could potentially refer to several concepts and we need to disambiguate: when somebody says 'Thyme goes well with rosemary,' we know that he did not say 'Time goes well with rosemary,' since the concept THYME[1] is much more related to ROSEMARY than TIME is.

[1] Throughout this thesis, we will use SMALL CAPS to denote concepts.

Given the importance of determining the semantic relatedness of concepts for human intelligence, it follows that this task is also significant for artificial intelligence—particularly for strong AI, i.e., when the goal is to build programs that have all mental capacities of humans. But it is helpful even from a weak AI perspective, i.e., when the goal is to develop intelligent applications tailored to specific problem domains, particularly in the realm of natural-language processing.

For instance, the Educational Testing Service (ETS) uses a measure of semantic relatedness in grading the essays every aspiring graduate student has to write [Landauer *et al.*, 1998]. In particular, they use a tool based on Latent Semantic Analysis (cf. Section 3.2) in order to compute the semantic similarity between an input essay and a set of training essays that have been graded by experts beforehands. In information retrieval, the same technique is mostly referred to as Latent Semantic Indexing [Manning *et al.*, 2008] and can be used to find documents that are semantically close to a query, even if there is no literal overlap.

Measures of semantic relatedness can also be utilized for spelling correction [Budanitsky and Hirst, 2001]. The rationale behind this application is that words spatially close in a natural-language text are often semantically close as well; so if a given word is unrelated to its surrounding context, while its spelling differs only slightly from another word in the dictionary that is closely related to the context, then the two words are likely to have been mixed up in a spelling error. The pair 'time' vs. 'thyme' in the context of 'rosemary' is an instance thereof.

As a last example, we name the problem of website optimization. In this context, semantic relatedness is leveraged to predict where on a given website users will click when trying to accomplish a given retrieval task [Kaur and Hornof, 2005]. This information can help Web designers streamline the appearance and organization of large website projects.

Since the notion of semantic relatedness is essential for many artificial intelligence applications, it is desirable to have algorithms that can infer this kind of knowledge automatically from data. Developing such algorithms is the core problem this thesis is concerned with. Thanks to the Internet, substantial amounts of useful data from which the algorithms could learn are readily available. In practice, however, a large part of the Web is not easily amenable to deep automated analysis, for several reasons: First, the Web is highly decentralized and consequently very diverse in terms of how information is encoded. It is estimated that 99.8% of the Web's content is hidden 'behind the query forms of searchable databases' [He *et al.*, 2007] and therefore cannot be systematically indexed by general search engines. Even in the accessible part of the Web, data formats are inconsistent, natural-language text being interleaved with bulleted lists, forms, images, advertisements, hyperlinks, etc. One may restrict oneself to pages containing only plain text, and still ungrammaticalities, slang, and typos will abound. Even for orthographically and grammatically correct natural-language text, parsing is an active research area, and even if we assumed parsing to be easy, natural-language *understanding* would still be a very hard task because human language relies so heavily on common-sense knowledge, which computers do not possess yet. In particular, natural-language understanding seems to require the notion of semantic relatedness as a 'subroutine', for instance for disambiguation, so we would face a chicken-and-egg scenario if we were to make the 'Wild Web' machine-understandable in order to infer semantic relatedness.

Consequently, most approaches have restricted themselves to shallower types of analysis. For instance, the aforementioned Latent Semantic Analysis [Landauer and Dumais, 1997] leverages co-occurrence statistics from a large corpus to embed words in a high-dimensional 'semantic vector space', while Pointwise Mutual Information using Information Retrieval [Turney, 2001] employs such statistics in an

information-theoretic way. These and similar shallow statistical approaches have the advantage of being easy to implement but also suffer from some common limitations: they cannot distinguish between types of relatedness (e.g., 'is-a' vs. 'is-part-of'); they cannot handle homonymy, i.e., disambiguate if a word has several senses; and they cannot deal with synonymy, i.e., they treat different words designating the same concept as distinct entities [Kaur and Hornof, 2005].

One way that could eventually enable a deeper understanding of Web content is by initially focusing on highly structured sub-Webs. Wu and Weld [2007], for instance, propose Wikipedia as a bootstrapping data set: the 'infobox' templates found there have slots for 'facts and statistics that are common to related articles' [Wikipedia, 2010a] and encode information in a structured and fairly unambiguous format. Therefore, they offer a body of basic knowledge that could be harnessed to extract some additional knowledge from the 'Wild Web' beyond Wikipedia, and so on in a positive feedback loop, every iteration making ever larger portions of the Web machine-understandable.

But even omitting Wikipedia's textual and infobox content, its raw hyperlink structure carries a significant amount of interesting information. The hypertextual Wikipedia graph alone can be viewed as a very primitive semantic network: articles represent the concepts, while the hyperlinks an article contains should constitute 'relevant connections to the subject of another article that will help readers to understand the current article more fully,' according to the Wikipedia linking guidelines [Wikipedia, 2010b]. In this sense, hyperlinks ideally represent semantic relationships between the concepts they connect.

It is this characteristic hyperlink structure which the work presented in this thesis builds on. We exploit it to infer semantic relatedness, and we enhance it by making it more complete. These contributions are summarized in more detail in the next section.

1.2 Contributions

We now state the four main contributions of this thesis.

The Wikispeedia Game. As discussed above, the Wikipedia link graph can be interpreted as a primitive semantic network. However, since it has not been developed with this explicit function in mind, it contains a lot of noise. We propose a human-computation approach that can mitigate this problem for the task of computing semantic relatedness: the online game Wikispeedia is played on Wikipedia and relies on the semantic value of most of its hyperlinks. We also give an analysis of typical game instances.

Computing Semantic Relatedness from Wikispeedia Data. The data collected through Wikispeedia do not directly contain numerical values of semantic relatedness. An additional computational step is necessary to extract such information. We present and evaluate an algorithm, grounded in information theory, which has this capacity.

Increasing Coverage through Dimensionality Reduction and Application to Wikipedia's Adjacency Matrix. In practice, the approach to inferring semantic relatedness from Wikispeedia data has the limitation of being undefined for many pairs of concepts. We propose algorithms based on principal component analysis (PCA) to increase coverage to the set of all pairs of Wikipedia concepts. These methods are general, i.e., not specifically tailored to Wikispeedia, and therefore applicable to other sparse measures of semantic relatedness as well. We demonstrate this by running our algorithms on the Wikipedia adjacency matrix.

Completing Wikipedia's Hyperlink Structure through Dimensionality Reduction. The hyperlinks connecting Wikipedia articles are crucial both for human usability and for artificial intelligence applications such as those outlined above. However, since they are added by humans, important links are often missing. We

propose an effective algorithm for automatically enriching the link structure, building on the dimensionality reduction techniques we also use for the task of inferring semantic relatedness.

1.3 Thesis Outline

The remainder of this thesis is structured as follows:

Chapter 2 describes the Wikispeedia game as well as similar projects and places them in the context of previous work in human-computation games.

Chapter 3 discusses Wikipedia as a resource for computing measures of semantic relatedness. We present a method for automatically analyzing data gathered through Wikispeedia in order to infer semantic relatedness, and investigate PCA as a means of making this measure more general. We argue that these techniques are applicable beyond the context of Wikispeedia and support this claim by demonstrating that they are able to compute semantic relatedness from Wikipedia's adjacency matrix.

In Chapter 4 we show that, beyond exploiting Wikipedia's hyperlink structure, our PCA-based method can also enhance it, by finding missing links, thus making it more coherent.

We conclude in Chapter 5 by recapitulating the main findings and discussing avenues for future research.

Section 3.3 is largely based on work presented at the *21st International Joint Conference on Artificial Intelligence* [West *et al.*, 2009a], while Chapter 4 is an updated version of a paper that appeared at the *18th ACM Conference on Information and Knowledge Management* [West *et al.*, 2009b].

CHAPTER 2

The Wikispeedia Game

This chapter introduces the online human-computation game Wikispeedia, which is played within Wikipedia and builds on the fact that the hyperlinks it contains are oftentimes predictable by common sense, since they usually bear semantic value. This game is at the heart of a large part of this thesis. In particular, we will later (in Chapter 3) show how it can be harnessed to define a measure of semantic distance, by effectively extracting the common sense human players have used during game play.

We proceed as follows: Section 2.1 defines the rules and shows screenshots of the game interface. Section 2.2 outlines Wikispeedia's history and provides usage statistics. In Section 2.3 we give details about the proof-of-concept implementation on which this thesis is based. Section 2.4 provides an overview of previous work in 'games with a purpose' and of game websites similar to Wikispeedia. We conclude in Section 2.5 by emphasizing typical characteristics of Wikispeedia games and by discussing Wikispeedia in the role of a 'game with a purpose'.

2.1 Rules

People play the game individually. To begin with, the player is given two Wikipedia articles. We refer to such a pair of articles as a *mission*. Starting from the first article, the goal is to reach the second one (the *goal article*), exclusively by following links in the articles encountered, minimizing the number of link clicks. Step-by-step backtracking is possible 'for free'.

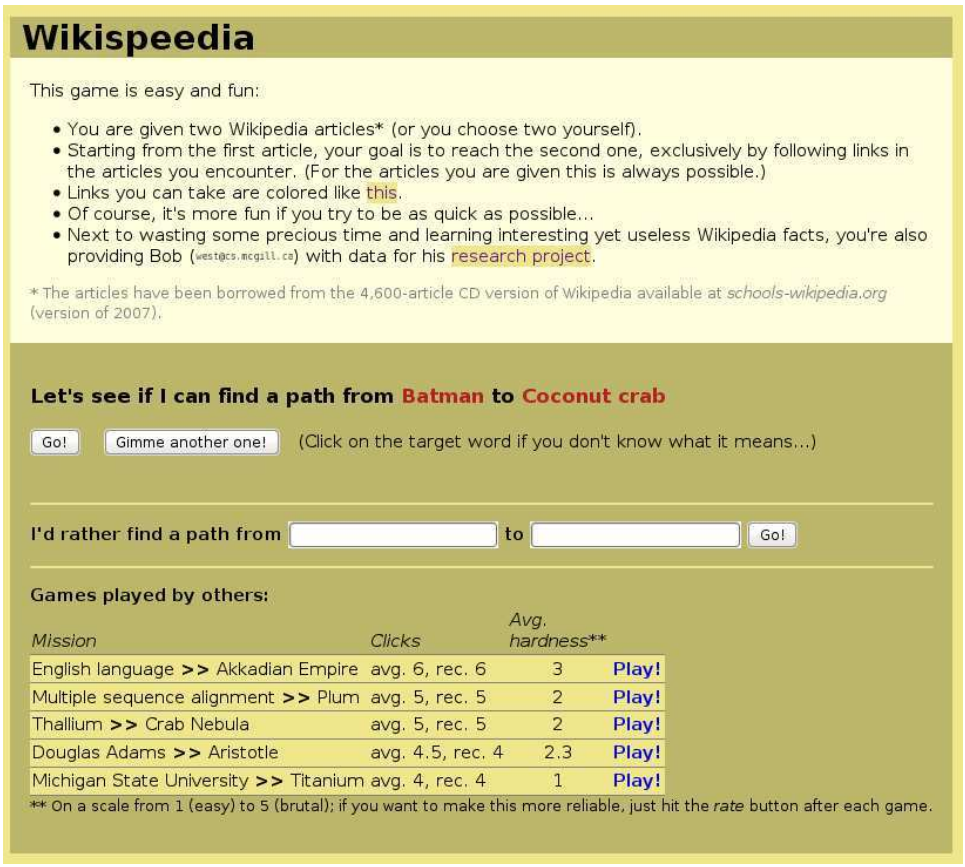


Figure 2–1: Screenshot of Wikispeedia's start page.

The interface is kept as simple as possible. It displays only the sequence of articles encountered so far, its length, and the content of the current Wikipedia article. Hyperlinks are highlighted for increased visibility. Figure 2–2 contains a sample screenshot.

After successfully completing a game, the player can optionally rate its difficulty and enter her name into the high-score table for the respective mission. If several players tie with respect to the number of clicks required, they are ranked according to the time they used. Competing with others for high-scores makes the game more attractive on a meta-level. A screenshot of the success page is shown in Figure 2–3.

Before the game starts, players can choose between three ways of obtaining a mission (see Figure 2–1 for a screenshot of the start page):

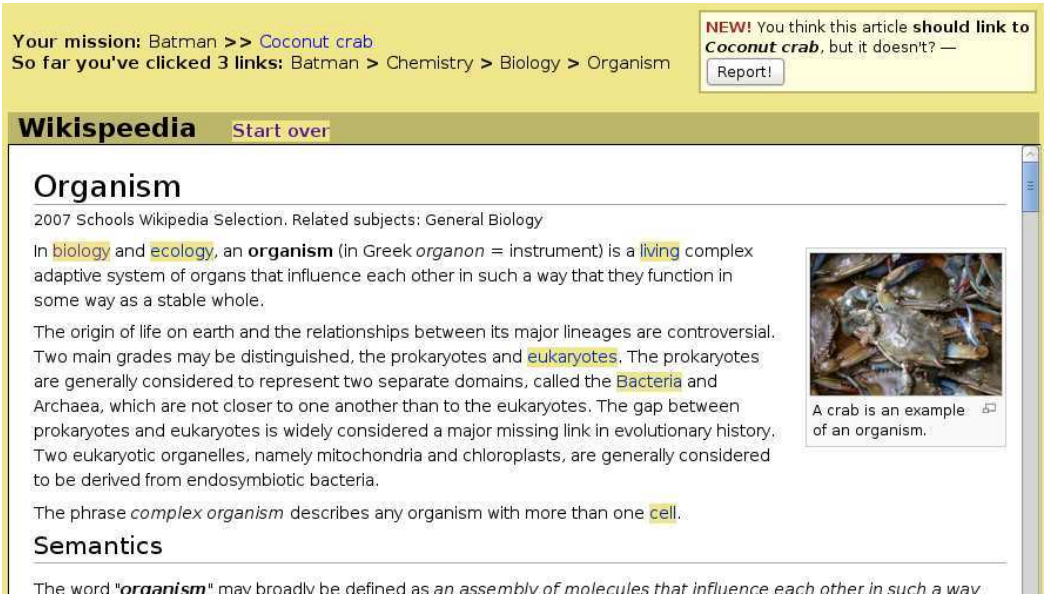


Figure 2–2: Screenshot of Wikispeedia. The box in the right upper corner is explained in Section 4.6.7.

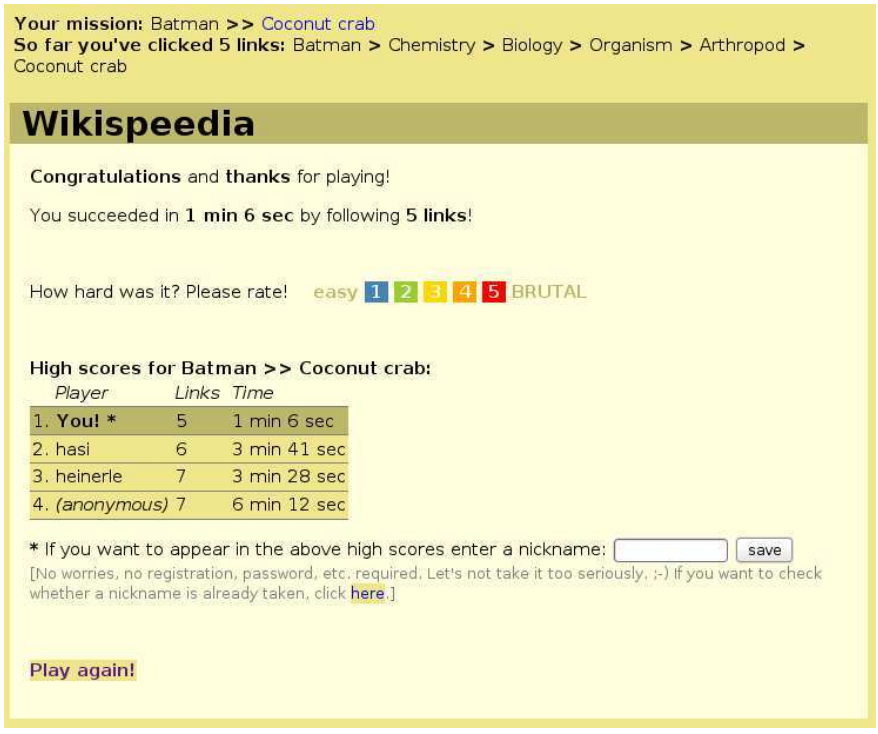


Figure 2–3: Screenshot of Wikispeedia's success page.

1. They can accept an automatic suggestion. (The choice of suggestions can be used to control the set of concepts about which data is collected.)
2. They can choose from a short list of missions that have been played by other players before. (The average and minimum numbers of clicks previous players needed is displayed as well, alongside the average difficulty rating, both of which can help the player make a choice.)
3. They can choose their own combination of any start and any goal article. (This is useful if the player wants to play a mission recommended by a friend, for instance.)

To lower the participation threshold, we decided to not require users to register with the website before being able to play. If they want to be listed in high-score tables, they simply pick an *ad-hoc* user name.

2.2 History

Wikispeedia is a version of the so-called Wiki Game [Wikipedia, 2010d] that has been played casually by Wikipedia users for a while. Although there do exist some alternative implementations of the Wiki Game (cf. Section 2.4.2), no analysis of data from this (or any similar) game has been done to date, to the best of our knowledge.

The Wikispeedia game website [West, 2009b] went online on August 11, 2008, and as of January 4, 2010, there have been 20,895 games. They originated from 6,482 distinct IP addresses from 80 countries. Table 2–1 shows the top 15 contributing countries. Clearly, the bulk of games were played in countries where English is either an official language or where people typically have a high level of proficiency in English (the Netherlands, Germany, Scandinavia), due to the fact that the game is played on an English version of Wikipedia.

United States	11,369	Germany	413	Sweden	141
Canada	2,508	New Zealand	260	Czech Republic	131
United Kingdom	2,046	Denmark	175	Ireland	127
Australia	873	Finland	152	Poland	126
Netherlands	743	Thailand	146	France	121

Table 2–1: Countries in which most games of Wikispeedia were played, as of January 4, 2010.

2.3 Proof-of-Concept Implementation

For the proof-of-concept implementation [West, 2009b] used for evaluating the approach proposed in this thesis it was important to have a clean set of Wikipedia articles about important concepts. Therefore, we chose the 2007 Wikipedia Selection for schools, which 'is a free, hand-checked, non-commercial selection from Wikipedia, targeted around the UK National Curriculum and useful for much of the English speaking world.' [Wikipedia, 2007] It is edited by SOS Children's Villages UK, fits on a DVD, and contains 4,604 articles that can serve as a free alternative to costly encyclopedias. As most Wikipedia articles are not present in it, the majority of links had to be removed, too. All links pointing to articles included in the collection were kept.

The game could be ported to full-size Wikipedia without a major effort. In our implementation, the articles are stored locally on the game website and the traces of players during games are stored in a database. No personal information is logged, except for IP addresses, which is necessary for estimating the number of distinct players.

2.4 Related Work

We now summarize previous work in human-computation games and present the alternative implementations of the Wiki Game we were able to identify.

2.4.1 Games with a Purpose

Collectively, humans spend vast amounts of time playing computer games. Luis von Ahn and colleagues have recently championed the idea of harnessing all those man-hours by designing games 'in which players perform a useful computation as a side effect of enjoyable game play' [von Ahn and Dabbish, 2008]. They call such games 'GWAPs', short for 'games with a purpose'.

The first and best-known example is the ESP Game [von Ahn and Dabbish, 2004], a version of which was later popularized under the name of Google Image Labeler [Google, 2010]. In this online game, two mutually anonymous players see the same image and enter words they associate with it. They both win as soon as they agree on a word. The fact that players cannot communicate with one another encourages them to supply common-sense labels that meaningfully describe the image. Since the players are independent, the labels they agree upon can be considered reliable.

Another game, Verbosity [von Ahn *et al.*, 2006], is more related to our work: it is inspired by the popular Taboo game and aims at collecting common-sense facts of the type 'A BICYCLE has WHEELS'.

In the article *Designing Games with a Purpose,* von Ahn and Dabbish [2008] later turned their experiences into a set of general design principles for GWAPs. It is of crucial importance that the game be fun in its own right, i.e., even if players have no interest in contributing to the project for which data is being harvested. This distinguishes GWAPs from other projects that have striven to collect data from volunteers over the Internet, such as Open Mind Common Sense [Singh *et al.*, 2002], which in general do not offer such an incentive.

Abstracting from the games they had designed previously, von Ahn and Dabbish [2008] identify three generic templates for GWAPs. Foregoing further detail, we note that all templates result in games played by two people in a collaborative

fashion: as in the concrete case of the ESP Game, players gain points if they agree with each other in one way or another, depending on the template.

Additionally, von Ahn and Dabbish provide a set of game features that can potentially be incorporated into instances of the aforementioned templates in order to make game play more enjoyable. In general, they aim at increasing challenge, an important aspect of successful games. Concretely, they mention the following features:

- Timed response: players have to accomplish a task in a limited amount of time.
- Score keeping: players get feedback on their performance, e.g., by earning points when they beat a previous record.
- Player skill levels: by playing more, people can gradually be promoted, e.g., from 'newbie' to 'grandmaster' status.
- High-score lists: scores are accumulated over time and displayed to players, to increase competition.
- Randomness: by selecting game instances at random, the difficulty level varies.

2.4.2 Alternative Implementations of the Wiki Game

We will now review, from a GWAP perspective, the three alternative implementations of the Wiki Game we have been able to identify. They have been developed independently of Wikispeedia, and, as far as we know, of each other. As opposed to Wikispeedia, which uses a condensed Wikipedia edition, they are all played on local copies of full Wikipedia. Also, these versions were designed as games only, without the human-computation component that Wikispeedia exhibits.

Wikirace [2010] is very similar to Wikispeedia. One difference is that there is a single persistent 'leader board', which ranks players based on a point system

that attributes credits for finishing a game, breaking a previous record, and creating a new mission (such new missions have to be approved by the manager of the website). On the contrary, Wikispeedia has a separate high-score table for each mission. In the terminology of Section 2.4.1, Wikirace has the features of score keeping, high-score lists, and randomness.

Wikipedia Maze [2010], too, has a scoring system in which points are gained for finishing and creating missions. This implementation also adds the option to vote up, or down, missions created by others. Whenever a mission a user created is voted up, she earns points; when it is voted down, she loses points. Missions can be tagged with keywords, such that users can actively select those that are likely to lead them through Wikipedia articles of interest to them. Finally, players get so-called 'badges' for achievements like creating a puzzle with 50 or more votes or having been an active member for more than a year. In summary, Wikipedia Maze has the features of score keeping, player skill levels (through the 'badges'), and high-score lists. With its voting and tagging systems, this implementation also adds a social dimension.

Wikipedia Game [2010] differs most from Wikispeedia, since it constitutes the only multi-player version of the Wiki Game, to the best of our knowledge. Players have 150 seconds to accomplish a predetermined mission and, doing so, race against each other in real time. They can communicate via chat and see how many clicks their competitors have made so far. They are notified once one of them has reached the goal article. Consequently, the primary objective in this implementation is to minimize time rather than the number of clicks. Wikipedia Game has the features of timed response, score keeping (via updates about the other players' status), high-score lists, and randomness.

2.5 Discussion

We now analyze typical characteristics shared by most game instances of Wikispeedia and discuss it as a 'game with a purpose'.

2.5.1 Typical Game Characteristics

There is a crucial difference between the way a computer and a human would play Wikispeedia, or any version of the Wiki Game. A computer would simply find the shortest path between the start and the goal, by any standard algorithm. This is clearly impractical for most humans. (A cheater could code a shortest-path finder, but we ignore this problem for now.) A human player will instead leverage semantic associations based on background knowledge of many common-sense facts, and select links according to this knowledge. Consider, for instance, the task of finding a path from SEYCHELLES to GREAT LAKES. It was solved in an actual game instance as follows:

⟨SEYCHELLES, FISHING, NORTH AMERICA, CANADA, GREAT LAKES⟩

This example showcases the anatomy of a typical game. Players try to reach, as quickly as possible, a general concept (in this case NORTH AMERICA), whose article has a lot of outgoing links. From such *hubs* it is easy to reach many parts of the Wikipedia graph. After this initial *'getting-away'* phase, the *'homing-in'* phase starts: the search narrows down again towards more specific articles that get more and more related to the goal.

Note the difference between the human path and the result of a shortest-path algorithm for this example: SEYCHELLES and GREAT LAKES are optimally connected by

⟨SEYCHELLES, ASIA, AMERICAN ENGLISH, GREAT LAKES⟩,

which is far less semantically meaningful than the path found by the human. In general, humans find intuitive, not shortest, paths. This observation is corroborated

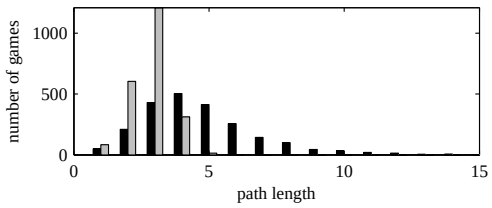


Figure 2–4: *Black:* histogram of lengths of 1,694 games (selected as described on page 33); the tail continues up to 30. *Gray:* histogram of shortest-path solutions to the same games.

by Figure 2–4, which shows that the distribution of game path lengths is shifted towards longer paths, compared to a shortest-path algorithm, and that it has a heavy tail towards longer paths. The median game of Wikispeedia consists of four clicks, while the median shortest path for those games consists of only three clicks.

In the next chapter we will show how the characteristic nature of Wikispeedia game paths can be exploited in order to define a measure of semantic distance between concepts.

2.5.2 Wikispeedia as a Game with a Purpose

It is important to note that, although we discussed the above implementations in the context of GWAPs, none of them is in fact a game with a purpose, since they all focus on being enjoyable, without leveraging the computation humans perform during game play. To the best of our knowledge, Wikispeedia is the first GWAP to employ Wikipedia as its 'playground'.

Wikispeedia is rather different from the basic game layouts described by von Ahn and Dabbish [2008] in that it is a competitive single-player game, whereas they only describe collaborative two-player templates. In a two-person game in which it is the players' goal to agree with one another, the results are likely to be of high quality, since two mutually anonymous persons produced them independently. Also, if the game has been designed appropriately, the output can be used as is (e.g., Verbosity [von Ahn *et al.*, 2006] directly produces common-sense facts). Extracting

useful information from raw Wikispeedia game traces is less straightforward, and consequently there is a significant computational intermediate step involved; this step represents one of the main contributions of this thesis.

According to von Ahn and Dabbish [2008], verifying the correctness of the game design is necessary even for collaborative two-player games, e.g., by paying independent humans to rate the quality of a sample of the data produced by players. We followed the same approach for Wikispeedia and were able to show that our computational analysis of raw Wikispeedia data does indeed result in valid semantic relatedness values (cf. Section 3.3.3).

Using the terminology of Section 2.4.1, Wikispeedia offers the features of score keeping and high-score lists (the number of clicks done at any point is displayed, and after completing a mission, players see a table showing how others have scored on the same mission previously), as well as randomness.

There is also an interesting and useful two-player version of Wikispeedia that could be envisioned, which, however, we did not implement: The two players would collaborate on one and the same mission, taking turns in clicking links. Keeping in mind that the next click would always have to be chosen by the partner, players would be encouraged to take more common-sense steps rather than random shortcuts that would confuse the partner. As a side effect, the gathered data would probably also reflect human common sense more accurately.

CHAPTER 3

Computing Semantic Relatedness Using Wikipedia

In Chapter 1 we have argued that, even if we ignore all textual content, we can still find a plethora of semantic information in Wikipedia's bare hyperlink structure, and Chapter 2 has described Wikispeedia, a game that capitalizes on the semantic content of Wikipedia links by using the hyperlink graph as its 'playground'. People make heavy use of their common-sense knowledge when playing Wikispeedia, as exemplified in Section 2.5.1. It is therefore desirable to develop methods able to extract this knowledge from recorded game traces. In this chapter we present such algorithms. They leverage Wikispeedia data in order to compute semantic distance. We also show that some of the same techniques can be applied directly to the adjacency matrix which captures Wikipedia's hyperlink structure.

We proceed as follows: In Section 3.1 we distinguish the notions of similarity and relatedness. In Section 3.2 we discuss related work. Section 3.3 introduces and evaluates Wikispeedia distance, while in Section 3.4 we investigate the use of dimensionality reduction for increasing its coverage. In Section 3.5 we apply the same dimensionality reduction techniques to Wikipedia's adjacency matrix. Finally, we recapitulate with a discussion in Section 3.6.

3.1 Similarity, Relatedness, and Distance

Many computational measures of semantic relatedness have been proposed, and in Section 3.2 we will review several of them. However, according to Resnik [1999], such approaches 'are seldom accompanied by an independent characterization of

the phenomenon they are measuring'. Instead, emphasis is put on the capability to emulate human performance in assessing semantic relatedness. Human performance is in turn explained by Quillian [1968] in terms of spreading activation in an associative semantic memory encoded in the brain.

Therefore, we do not attempt to give a formal definition of semantic relatedness either but rather treat it as an empirically defined notion: two concepts are considered related if a majority of human respondents say they are; this might be the case, e.g., because the concepts often co-occur in similar everyday situations, because one concept is a more specific version of the other, or because one concept is the opposite of the other.

In the cognitive science community, similarity and relatedness are often taken to constitute distinct notions, *similarity* being a special case of *relatedness*. Resnik [1999] gives the following example: CARS and GASOLINE are more related than CARS and BICYCLES; nonetheless, CARS are more similar to BICYCLES than they are to GASOLINE.

Similarity, being more specific, has been subjected to a more formal definition than relatedness. Psychologist Amos Tversky [1977], for instance, defines similarity as the result of a feature-matching process. Every entity has a vector of features, and measuring the semantic similarity between two entities amounts to computing the similarity, or overlap, between their feature vectors.

Tversky also argues that similarity—and hence relatedness and distance—is not necessarily symmetric (and he adapts his model accordingly). For instance, he found empirically that humans consider POLAND to be more similar to the USSR than they consider the USSR to be similar to POLAND. Therefore, semantic distance does not comply with the geometric notion of distance, which calls for symmetry.

A further difference is that semantic distance does not satisfy the triangle inequality, which Tversky calls 'hardly compelling' in the context of semantic distance, giving this counterexample: JAMAICA is highly similar to CUBA, both being islands in the Caribbean Sea, and CUBA is highly similar to the USSR, both being communist countries. According to the triangle inequality, JAMAICA would be very similar to the USSR, too (their distance would be at most the sum of the other two small distances). Notwithstanding, JAMAICA is highly dissimilar to the USSR.

Tversky's features encompass 'appearance, function, relation to other objects, and any other property of the object that can be deduced from our general knowledge of the world.' [Tversky, 1977] Current computers do not possess such general knowledge yet, so, instead, most computational methods devised to date for inferring semantic relatedness or similarity have drawn on the statistics of text corpora or the structure of semantic network graphs. We will discuss these approaches next, with a bias towards work that involves Wikipedia.

The term *semantic distance* can be used to designate the opposite of either semantic similarity or semantic relatedness. The semantic distance measures we introduce in the remainder of this thesis are in terms of relatedness, since they do not adhere to the definition of similarity as just delineated.

3.2 Related Work

Researchers have developed numerous techniques for inferring the degree of relatedness between two concepts based on their relative position in a semantic network. For instance, Rada *et al.* [1989], inspired by Quillian's [1968] work (cf. Section 3.1), proposed a shortest-path metric, according to which the degree of similarity is determined by the length of the shortest path between two vertices in the semantic

network graph. Resnik [1999] defines the similarity of two concepts as the information content of their least common subsumer in the taxonomical hierarchy of WordNet [Fellbaum, 1998], a widely used semantic network.

This work dates back to the time before the launch of Wikipedia, which later proved to be an invaluable resource of semantic information. In their extensive review of the field of knowledge extraction from Wikipedia, Medelyan *et al.* [2009] even maintain that Wikipedia ushered in a 'new era of competition' in the domain of semantic relatedness.

For instance, Strube and Ponzetto [2006] explore the use of WordNet-based techniques, such as Rada *et al.*'s [1989] and Resnik's [1999], in the context of the Wikipedia graph. Since the latter is very densely connected, they use only the graph (approximately a tree) of Wikipedia categories rather than the entire hyperlink structure.

However, there do exist approaches that exploit Wikipedia's full hyperlink content. Ollivier and Senellart [2007] define a Markov chain on the link graph and use tools from the theory of random walks to find related pages.

Ollivier and Senellart compare their method, among others, to what they call 'cosine with tf-idf weight'. This is equivalent to Milne's [2007] out-link–based measure (although it is not recognized in either paper), which represents articles as vectors of outgoing links and computes their similarity using the cosine measure [Manning *et al.*, 2008]. Milne later augmented his algorithm by combining it with a metric based on incoming rather than outgoing links [Milne and Witten, 2008a], inspired by the Normalized Google Distance [Cilibrasi and Vitányi, 2007].

All of the above approaches are graph-based. Another important class of semantic relatedness measures are the vector-based ones. Their most important classical (i.e., pre-Wikipedia) representative is Latent Semantic Analysis (LSA) [Landauer and Dumais, 1997]. It constructs a term–document frequency matrix from

a large corpus and makes use of PCA to reduce its dimensionality, thus educing information that is present in the data only implicitly (hence the attribute 'latent'). Concepts are then compared using the cosine similarity between rows in the reduced matrix.

As is obvious in the name, Explicit Semantic Analysis (ESA) [Gabrilovich and Markovitch, 2007] is inspired by Latent Semantic Analysis. It profits from the fact that Wikipedia's content is highly structured, by defining meaningful dimensions of semantic space explicitly in terms of Wikipedia articles (whereas the principal components found by LSA do not necessarily have an intuitive meaning).

Veksler *et al.* [2008] construct an explicitly defined vector space as well. They propose a meta-algorithm: The term–term matrix they work with contains relatedness values originating from another, arbitrary measure of semantic relatedness. While any choice is admissible, they use the aforementioned Normalized Google Distance. In the next step, they reduce the dimensionality of this data matrix. While LSA uses PCA for this purpose, Veksler *et al.* do so by picking a subset of matrix columns, using a genetic algorithm.

An advantage of such vector-based models is that they can compute relatedness values for pairs of entire documents as opposed to just pairs of words.

For an extensive literature review about knowledge extraction from Wikipedia, beyond the task of inferring semantic relatedness, we refer the reader to Medelyan *et al.* [2009].

3.3 The Wikispeedia Method

In this section we explain how we leverage the semantic knowledge implicit in Wikispeedia game traces to derive a distance measure between concepts. We also describe some important properies of the resulting distance measure and evaluate it empirically.

3.3.1 Proposed Semantic Distance Measure

As we have argued in Chapter 1, Wikipedia's structure can be considered to be a rudimentary semantic net. Consequently, it seems reasonable to apply techniques such as Rada *et al.*'s [1989] shortest-path metric (cf. Section 3.2) to the Wikipedia link graph. But using the raw hyperlink structure of Wikipedia leads to several problems. First, while many hyperlinks correspond to semantic links, many others do not. Links are often added based on the inclination of the author, rather than because the concepts are related. Also, if one looks only at the presence or absence of links, no distinction can be made between closely and loosely related concepts. This leads to a combinatorial explosion, such that every page is connected to every other page by 4.6 links on average [Dolan, undated]. For instance, both BASEBALL and ARCHIMEDES have distance 2 to CARL FRIEDRICH GAUSS according to the shortest-path metric, although clearly the latter is relevant while the former is not.

Neither could a purely path length–based measure account for the frequency with which Wikispeedia players choose an article to reach a goal. But clearly, if many players pick a specific article, it should be considered more related to the goal than if only few do. This is why the semantic distance measure we propose is based on information theory. Intuitively, it quantifies how many bits are needed to encode a common-sense Wikipedia path between two concepts. The fewer bits are needed, the more strongly the two concepts are related. The number of bits required is smaller if the path complies with other paths connecting the same articles. In order to formalize this idea, we must first discuss *click probabilities*.

Click Probabilities. Let A, A', and G be random variables representing the current Wikipedia page, the next Wikipedia page, and the goal page of a game. For any Wikipedia article a and any Wikipedia goal (or target) article g, one can consider the probability distribution $P(A'|A = a, G = g)$ over a's out-links. This distribution is multinomial and specifies, for each article a' that can be reached in

one hop from a, the probability that a player continues to a' if she is currently on a and is trying to find goal article g. This can be estimated from the observed games using standard Bayesian methods, as the mean of the Dirichlet distribution which is the conjugate prior of $P(A'|A=a,G=g)$.[1] We use P^* to denote the *posterior click probability* estimated after seeing all the data:

$$P^*(A'=a'|A=a,G=g) = \frac{N(A'=a',A=a,G=g)+\alpha}{N(A=a,G=g)+\alpha\,\text{outdeg}(a)}, \tag{3.1}$$

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

Before observing any games (i.e., if all N-counts in (3.1) are zero) the estimate is the uniform *prior click probability:*

$$P^0(A'=a'|A=a,G=g) = 1/\,\text{outdeg}(a) \tag{3.2}$$

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

$$d_p(a_i,g) = \frac{-\sum_{j=i}^{n-1}\log P^*(A'=a_{j+1}|A=a_j,G=g)}{-\log\text{PageRank}(g)}. \tag{3.3}$$

In the numerator, $-\log P^*(A'=a_{j+1}|A=a_j,G=g)$ is the information content of the link from a_j to a_{j+1} given that the goal is g, or in other words, the number of bits needed to represent that link optimally in a Huffman coding. So the numerator

[1] This amounts to simply counting how often each link was clicked, smoothed by starting the counters with a value $\alpha>0$ instead of 0.

sums up the numbers of bits needed to code each separate link that was clicked along p, and consequently indicates the number of bits needed to code the entire path (note that this is conditional on g).

The denominator contains the PageRank [Brin and Page, 1998] of the goal article g, which is the stationary probability of g during a (fictional) random walk on the Wikipedia graph. We implemented the PageRank algorithm and ran it locally on the Wikipedia graph to get these numbers. One can think of PageRank(g) as the prior probability of being in article g, and of the entire denominator as g's information content, or the number of bits needed to code article g independently of any game. This serves the purpose of normalization: intuitively, a concept that is hard to reach (hard to 'explain') is allowed to be related to concepts that are farther from it on Wikipedia paths. For instance, UNITED STATES has PageRank 0.010 (1% of time steps on a random walk will be spent on the UNITED STATES article), while TURQUOISE has a PageRank of only 5.8×10^{-5}. Since $-\log(0.010) \approx 6.6$ and $-\log(5.8 \times 10^{-5}) \approx 14$ (about twice 6.6), a path p from an article a to goal TURQUOISE may take twice as many bits to code as a path q from some article b to goal UNITED STATES, and still we will have $d_p(a, \text{TURQUOISE}) \approx d_q(b, \text{UNITED STATES})$.

Instead of using uniform transition probabilities (cf. (3.2)) for the random walk, as in the standard PageRank algorithm, it might seem better to use the transition probabilities estimated from data (cf. (3.1)). Such a 'posterior PageRank' would indicate how hard it is to find an article while one is actively looking for it, rather than wandering aimlessly. Numerically, however, this is a minor difference, so the results we present here use the standard PageRank.

Path-independent Distance. So far, we have described distances that are derived from single paths. To get a *path-independent distance* from a to g, we simply average over all paths running through a and reaching goal g. Thus, if $\mathfrak{P}$ is

the set of such paths,

$$d(a,g) = \frac{1}{|\mathfrak{P}|} \sum_{p \in \mathfrak{P}} d_p(a,g). \tag{3.4}$$

If an article a never occurred in a game with goal g then $d(a,g)$ is undefined. Therefore, our method is incremental, with the number of article associations that are established growing as more game data is gathered. In Section 3.4 we will show how undefined entries can be eliminated by using generalization techniques.

Properties. With respect to the distinction we made in Section 3.1, our distance is in terms of relatedness, not similarity. Players try reaching the goal by passing through articles that are increasingly likely to link to it, which is usually the case because the concepts they represent are more and more related (but not necessarily more similar) to it. Hence, there is an intuitive interpretation of what $d(a,g)$ measures, along these lines: When one concentrates on the goal concept g, how much of a 'mental leap' is it for concept a to pop up in one's mind?

It follows that an important property of our proposed distance measure is that it is not symmetric: in general, $d(a,b) \neq d(b,a)$. Although it could be easily symmetrized (e.g., by taking $\min\{d(a,b), d(b,a)\}$), we do not do this, because asymmetry is desirable for psychological reasons (cf. Section 3.1). For instance, $d(\text{MINNEAPOLIS, MINNESOTA}) = 0.22$, while $d(\text{MINNESOTA, MINNEAPOLIS}) = 0.12$. Intuitively, this makes sense: when one thinks of MINNEAPOLIS, MINNESOTA is probably one of the first associations, because MINNEAPOLIS is in MINNESOTA. On the flip side, there are many other places in MINNESOTA one could think of, e.g., ST. PAUL, so when thinking of MINNESOTA, MINNEAPOLIS is not as predominant an association. We note that this asymmetry could also be helpful when labeling concept relationships with their type, which is often directional (e.g., 'is-part-of'). However, we do not address this issue here.

Unlike shortest paths, our measure also does not fulfill the triangle inequality: in general, $d(a,c) \not\leq d(a,b) + d(b,c)$. Section 3.1 exemplified why this, too, is an

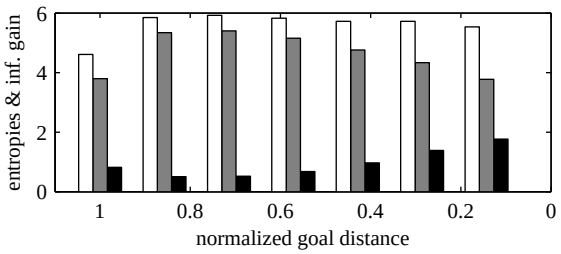


Figure 3–1: *White:* prior entropy. *Gray:* posterior entropy. *Black:* information gain. (Averaged over 1,694 games, selected as described on page 33.)

asset. Generally, the triangle inequality can be considered to model the transitivity of relatedness. It should be noted that our method can still capture transitive higher-order relatedness, but only when this is suggested by common sense, not by the structure of the graph: even if there is no direct link between two articles, the two will be considered related if people often went through one when aiming for the other.

Entropies. Having defined posterior and prior click probabilities (cf. (3.1) and (3.2)), we may now also revisit in more quantitative terms the observation of Section 2.5.1 that games typically consist of a getting-away phase followed by a homing-in phase. Consider Figure 3–1, which shows how the entropies of click probability distributions associated with articles vary along game paths. Since the path length varies among games, we normalized it to $[0,1]$, the *normalized goal distance* of the i-th article on a path consisting of n articles being $(n-i)/(n-1)$. For averaging over all games, we discretized $[0,1]$ into seven equally sized intervals and computed the means of three quantities for each interval. The left bar is the *prior entropy* H^0 of P^0. The middle bar is the *posterior entropy* H^* of P^*. Entropy measures the uncertainty associated with a distribution, so the right bar, $H^0 - H^*$, shows the loss of uncertainty afforded by seeing the recorded games. We call this quantity *information gain*.

First note that the prior entropy H^0 of the click probability distribution associated with an article a is simply $\log(\text{outdeg}(a))$, which is implied directly by (3.2) and the definition of entropy. The fact that this quantity, represented by the left bar, initially increases shows that there is indeed a getting-away phase in which players strive to reach a hub—an article with high out-degree.

Moreover, since all players share the same common sense, they perform many steps in similar ways; e.g., if NOAM CHOMSKY is the goal and a player is currently on LANGUAGE, she is much more likely to proceed to LINGUISTICS than to GORILLA. This is why the information gain is high at the start, as players get away to the same hubs, but decreases in the middle of the game; then the gain increases again, as they home in using the same common sense. In other words, the initial getting-away and the final homing-in phases are much more predictable after seeing game data than before. In the next section, we will explore the possibility of using information gain to split the getting-away from the homing-in phase, which can help in filtering out articles that are unrelated to the goal article of the respective path.

Data Set Preprocessing. On a technical note, we remark that we preprocess the data set of game traces in two ways.

First, we discard spurious game paths of excessive length, choosing twice the length of the longest shortest path in the entire link graph (which is typically much larger than the median shortest path) as the maximum allowable threshold. Only 1.9% of all paths fall into this category. The eliminated paths are most probably the result of aimless random walks and can be considered noise.

Second, we inflate the data set for computing the posterior click probability (cf. (3.1)) by including every path twice, once in its original form and once with the last article removed, which amounts to making the penultimate article the goal article. Eye-balling the recorded games revealed that such crippled paths still 'make sense',

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

Table 3–1: Concepts a and $d(a, \text{NOAM CHOMSKY})$. Canceled entries are those eliminated by the method of Section 3.3.2.

i.e., follow the same dichotomy—getting away vs. homing in—as actual paths. We did not evaluate this heuristic formally.

Example. To illustrate our distance measure, we provide an example. Table 3–1 shows all concepts with a defined distance to NOAM CHOMSKY, in order of increasing distance, i.e., decreasing relatedness. Note that the data comes from only nine games with goal NOAM CHOMSKY.

3.3.2 Filtering Unrelated Concepts

Subjectively, Table 3–1 seems reasonable. The top six concepts are all highly related to NOAM CHOMSKY. However, further down the list we have a mix of related and unrelated concepts. One would like to discriminate automatically which of these associations are truly meaningful. We have seen that the typical anatomy of games is 'get away to hub, then home in on goal'. Since we compute distances to the goal for *all* articles along the path, the articles from the getting-away phase get defined distances to NOAM CHOMSKY, too. However, typically they are no more related to the goal than the hundreds of other concepts whose distances to NOAM

CHOMSKY are undefined simply because they never occurred in games with that goal. So, in order to eliminate irrelevant entries, our approach should exclude the articles of the getting-away phase when computing distances. We experimented with three methods for predicting the start of the homing-in phase (results are reported in Section 3.3.3):

1. The shape of information gain in Figure 3–1 suggests that we might, for a given game instance, assume the homing-in phase to begin with the article where the information gain starts to increase again.
2. Alternatively, one can collect ground-truth examples of the *split position* (the article that starts the homing-in phase) and split all paths at the most likely (i.e., average) split position computed from the labeled examples.
3. A third method can be obtained by combining both information gain and position along the path as the features of a machine learning classifier that predicts the start of the homing-in phase.

In any case, an article would be erased from lists such as Table 3–1 if it never occurred in the homing-in phase of a game with the given goal. We eventually opted for the second of the methods just sketched, which is justified by the evaluation we present in the next section.

3.3.3 Results

The data used in this evaluation comes from the implementation described in Section 2.3. In order to gather data that is useful for our purposes, the same goal article was specified in multiple automatic game suggestions (cf. Section 2.1) and was thus played by different players. Initial articles, however, were often chosen at random. We let $\alpha = 0.1$ in (3.1).

We conducted the evaluations on Amazon Mechanical Turk [Amazon, 2009], an online platform on which 'requesters' can post questionnaires (among many

Figure 3–2: Amazon Mechanical Turk task for learning how to split game paths into the getting-away and homing-in phases. Each task consisted of five lists of the depicted type.

other types of tasks), which are subsequently completed for a typically small amount of money by 'workers', regular Internet users who have registered with the system. It has been shown that non-expert labels obtained through Mechanical Turk agree very well with gold-standard expert annotations for natural-language tasks [Snow *et al.*, 2008], which justifies using it for our purpose.

Filtering Unrelated Concepts

Using Amazon Mechanical Turk [Amazon, 2009], we had human raters mark the split position of 500 game paths. Every path was split by two different people, such that we can gauge inter-human agreement. In each task, five paths had to be split. The task instructions were as follows, with 'red target word' referring to the title of the goal article of the respective game (see Figure 3–2 for an example task):

> 'Below, you are seeing five lists of words. Each list starts with words that are generally not related to the red target word, and it ends with words that are highly related to the red target word. For each list, please do the following:
>
> Split each list in two, i.e., mark exactly one checkbox in each list, such that all the words to the right of that checkbox are highly related to the red target word, whereas not all the words to the left are.'

Experimenting with the hand-coded rule that splits the path after the article with minimum information gain (the first method in the list of Section 3.3.2), we obtained better results when we restricted the potential split positions to the second

half of the game path. Using cross-validation, we found that this hand-coded heuristic is on average 0.91 positions off the actual split position as defined by humans on Mechanical Turk.

Humans split paths on average at a normalized goal distance of 0.40, i.e., slightly after the middle of a game. A predictor that simply starts the homing-in phase with the first article that has a normalized goal distance less than or equal to 0.40 (the second method of Section 3.3.2) has an average offset of 0.78 positions from the actual split position. So this predictor is relatively better than the one based on information gain. In absolute terms, too, this result is good, since the average game in the labeled data set consists of as many as 5.7 articles.

We also tried combining both features in a neural net classifier (the third method of Section 3.3.2). The network has one hidden layer (two units) and two input features: the number of links between the input article and the article with minimum information gain in the second half of the path, and the normalized goal distance of the input article. The class label was 1 if the input article was the one labeled by the human, and 0 otherwise. Once the network is trained, we use it to split unseen paths as follows. For every article along the path, we feed its two features into the network and compute the network output. We predict the relevant part of the path to start with the article for which the net outputs the highest value. This predictor is on average 0.77 positions off the actual split position, i.e., it is slightly better than the simple predictor based solely on normalized goal distance. However, we do not consider the gain of 0.01 to be large enough to outweigh the better computational efficiency of that simpler predictor, which we therefore use in the remainder of this thesis to split paths into getting-away versus homing-in phase. As mentioned in Section 3.3.2, we subsequently exclude the articles of the getting-away phase when computing semantic distance from game paths.

Human Evaluation of the Distance Measure

In order to test the quality and psychological validity of our distance measure, we compare it to Latent Semantic Analysis (LSA; cf. Section 3.2). We chose LSA because (1) it seems to be the method most widely applied to real-world problems, e.g., automated essay grading [Landauer *et al.*, 1998], (2) it is readily available via a Web interface [Landauer and Kintsch, 1998], and (3) it has been cognitively validated by the psychological community, not only in psychometric but also in behavioral experiments [Huettig *et al.*, 2006].

Since the method described in Section 3.3.1 is incremental, defining a distance only for pairs that co-occurred in at least one game, we cannot compare to a standard test set of human-labeled concept pairs (e.g., WordSimilarity-353 [Finkelstein *et al.*, 2002]), since there would be too little overlap between the pairs covered by our method and the test set. Instead, we resorted to querying humans directly, as follows.

The data set evaluated contained 1,694 games, collected from players with 282 distinct IP addresses. The set of goals was constrained to 124 randomly selected articles. Each of these 124 target concepts was the goal of between 7 and 26 (median 12) games. For each target, the five closest semantic neighbors were picked according to our method and the LSA method, respectively. For LSA, we used the same corpus as Huettig *et al.* [2006]: 'General Reading up to 1st year college' (300 factors). Since we wanted to test for semantic (rather than merely phonetic) relatedness, we did not consider as neighbors words containing the target word or contained in it (e.g., CHOMSKY and NOAM CHOMSKY), the plural of the target, and adjectives directly derived from the target (e.g., CHINA and CHINESE). This yielded usually a set of ten neighbors for each concept. If both methods agreed on a word, it was included just once, and the neighbor set contained only nine concepts (this happened for eleven targets). If they agreed on two words, the set contained eight

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

Table 3–2: Results of the comparison of the Wikispeedia method to LSA.

concepts (this happened for three targets). Larger agreements were not encountered. For each target concept, four different human raters were given the neighbor set on Amazon Mechanical Turk (the order of entries in the set was randomized) and asked to select the three words they considered most closely related to the target.

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

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

while our method can handle multi-word concepts (Wikipedia article names may contain several words, e.g., WORLD HEALTH ORGANIZATION).

3.4 Increasing Coverage through Dimensionality Reduction

In the previous section we demonstrated that, given a specific target concept, the top matches of our distance measure comply with the human notion of semantic relatedness significantly better than the top matches of LSA. In other words, our distance measure has very high precision, i.e., when a pair of concepts gets a low distance value, the concepts are most often highly related in reality, too. However, to stick to the terminology of information retrieval, recall is rather low, i.e., many of the concepts that are in fact highly related to a given concept do not get a low distance value. Rather, the respective distances are undefined, due to the incremental nature of the technique. Thus, the method is slow in comparison to corpus-based methods, even though data collection is facilitated by the fact that it is enjoyable for human contributors.

In this section we investigate different ways to overcome this low coverage, by generalizing Wikispeedia distance to concept pairs whose constituents never co-occurred in the same game. The approaches are based on dimensionality reduction, more specifically on principal component analysis (PCA) [Pearson, 1901].

3.4.1 Transforming Distance into Relatedness

Let N be the number of concepts, i.e., of Wikipedia articles. Then $\mathbf{D} = \left(d_{ij}\right)$ is the $N \times N$ square Wikispeedia distance matrix, where entry d_{ij} equals the distance $d(i, j)$ between concepts i and j,[2] as computed from Wikispeedia games.

[2] For convenience, we refer to an article by its name and by its index in the (say, alphabetical) list of articles interchangeably.

As mentioned, this quantity is undefined for most concept pairs (including those filtered by the method of Section 3.3.2). However, the approach we take here involves numerical matrix operations, which require that the matrix be defined everywhere. So we consider previously undefined distances to have maximum value. Subsequently, we transform distance into relatedness, since the benchmark data set to which we compare measures the latter, not the former. Let $d_{\max} = \max_{i,j}\{d_{ij}\}$ be the maximum entry of $\mathbf{D}$; then the *Wikispeedia relatedness matrix* is defined as

$$\mathbf{W} = \left(w_{ij}\right) = \left(1 - \frac{d_{ij}}{d_{\max}}\right). \tag{3.5}$$

The transformation of (3.5) also normalizes all values by mapping them into the interval $[0,1]$. Entries that were originally undefined in $\mathbf{D}$ are 0 in $\mathbf{W}$, and as a result, $\mathbf{W}$ is very sparse.

Note that, since the original $\mathbf{D}$ is asymmetric, $\mathbf{W}$ is asymmetric, too, i.e., in general, $w_{ij} \neq w_{ji}$. To make this distinction clear, we will say the row indices of $\mathbf{W}$ refer to *source concepts,* while the column indices refer to *target concepts*.

In the next section we apply PCA to $\mathbf{W}$, for which it is a technical requirement that the matrix be centered around the mean, by subtracting the respective column mean from each column. Note that this operation makes the matrix much less sparse by mapping many zeros to negative values. In what follows, we assume $\mathbf{W}$ to be mean-centered, unless noted otherwise.

3.4.2 First-Order Method

One way to reduce the sparse coverage of the Wikispeedia relatedness measure in a meaningful way is to smooth the relatedness matrix $\mathbf{W}$ using PCA. We may regard $\mathbf{W}$ as a data matrix as typically used in PCA. Row vectors represent data points, consisting of the values of different features for that data point. The data points are source concepts, and the features are the relatedness values to all the

target concepts. Thus, a source concept is represented by its relatedness to the target concepts. The source concepts form a cloud of points in an N-dimensional vector space (let us call it *concept space*). After mean-centering, the *average source concept* sits in the origin.

Eigenconcepts and Eigenspace. This point cloud is not uniformly distributed but rather sprawling in certain directions and squished in others. This is due to correlations among the points: source concepts that are all related to one specific target concept often share a relatedness to other particular target concepts as well. For example, concepts related to ADAM will often be related to EVE as well. PCA finds the directions along which the point cloud is spread out most, i.e., along which source concepts tend to differ most from the average source concept. Those directions are called *principal components*. They are vectors in the N-dimensional concept space pointing away from the average source concept in the origin; by convention, they are normalized to a length of 1.

The principal components found are orthogonal. Hence, an appealing geometric way of thinking about PCA is as a rotation of the axes of the co-ordinate system such that the spread (more formally, the variance) of the data is k-th largest along dimension k; it then computes the co-ordinates of each point in the new basis formed by the principal components. The principal components themselves can be considered 'synthetic' source concepts (since they are points in the N-dimensional concept space).

Mathematically, the principal components are the eigenvectors of the data covariance matrix. Hence, we call them *eigenconcepts,* to emphasize that they are eigenvectors and points in concept space. The new space resulting from the rotation is called *eigenspace*.

Eigenspace Projection. Computing the co-ordinates of a source concept $\mathbf{w}_i$ (a row vector of $\mathbf{W}$) in eigenspace amounts to projecting it onto the eigenspace

basis vectors, i.e., onto the eigenconcepts $\mathbf{e}_k$. Then the vector $\mathbf{p}_i$ of projections is the eigenspace representation of $\mathbf{w}_i$:

$$\mathbf{p}_i = (p_{i1}, ..., p_{iN}) = (\mathbf{w}_i \mathbf{e}_1^{\mathrm{T}}, ..., \mathbf{w}_i \mathbf{e}_N^{\mathrm{T}}) \tag{3.6}$$

In matrix notation this can be written succinctly as

$$\mathbf{P} = \mathbf{W}\mathbf{E}^{\mathrm{T}}, \tag{3.7}$$

where projection vector $\mathbf{p}_i$ is the i-th row of $\mathbf{P}$ and eigenconcept $\mathbf{e}_k$ is the k-th row of $\mathbf{E}$.

Since PCA performs a rotation, $\mathbf{E}^{\mathrm{T}}$ is a rotation matrix, i.e., $\mathbf{E}^{\mathrm{T}}\mathbf{E}$ is the identity matrix. Thus, the reverse projection from eigenspace back into concept space (the so-called *reconstruction* of $\mathbf{W}$) is

$$\mathbf{W} = \mathbf{P}\mathbf{E}. \tag{3.8}$$

Expanding (3.8), a single entry of $\mathbf{W}$ is computed in the reconstruction as follows:

$$w_{ij} = \sum_{k=1}^{N} p_{ik} e_{kj} \tag{3.9}$$

Entry w_{ij} is large when there are many eigenconcepts $\mathbf{e}_k$ that (a) are important components of $\mathbf{w}_i$ in eigenspace (resulting in large p_{ik}) and that (b) are themselves related to target concept j (resulting in large e_{kj}). Remember that eigenconcepts live in concept space and have 'synthetic' relatedness values to 'real' target concepts, e_{kj} being the k-th eigenconcept's relatedness to target concept j.

Dimensionality Reduction. The reconstruction of $\mathbf{W}$ as $\mathbf{P}\mathbf{E} = \mathbf{W}\mathbf{E}^{\mathrm{T}}\mathbf{E}$ is exact. However, getting an exact reconstruction is not useful from our point of view. We want the reconstructed matrix to be smoothed, i.e., we want entries that were left undefined by the Wikispeedia method and are thus equal to 0 in $\mathbf{W}$ (before mean-centering) to be increased in a meaningful way where it is justified. This is

an important difference compared to more traditional applications of PCA, in which one wants to obtain a reconstruction that is as exact as possible. We actually want to obtain a reconstruction that enriches the original data. For instance, imagine we want to enrich in this way source concept i, represented by the i-th row vector $\mathbf{w}_i$ of $\mathbf{W}$. What we would like, intuitively, is to find first a set $\mathfrak{C}$ of source concepts similar to $\mathbf{w}_i$ (similar in a vector sense, that is). Then, if i's relatedness to j is undefined according to the Wikispeedia measure but many of the vectors in $\mathfrak{C}$ indicate a high relatedness to j, then, by analogy, i should be considered highly related to j as well. This reasoning scheme has been called *cumulative analogy* [Chklovski, 2003; Speer *et al.*, 2008].

In order to augment the data this way, we must first ensure that we 'forget' some information while dwelling in eigenspace, just like in the case of traditional dimensionality reduction. First, we project a source concept $\mathbf{w}_i$ from concept space into eigenspace, obtaining its eigenspace representation $\mathbf{p}_i$ (cf. (3.6)). Now we 'shrink' $\mathbf{p}_i$ by setting to zero all components p_{ik} with $k > K$, for some fixed K. These were the projections onto eigenconcepts along whose direction the variation in the data is small, so it can be considered noise. By shrinking $\mathbf{p}_i$ we eliminate that noise. Now we can reconstruct $\mathbf{w}_i$ approximately by projecting it back into concept space (cf. (3.8)):

$$\mathbf{W}_K = \mathbf{P}_K \mathbf{E}_K, \tag{3.10}$$

where $\mathbf{P}_K$ consists of the first K columns of $\mathbf{P}$ and $\mathbf{E}_K$ of the first K rows of $\mathbf{E}$. Matrix $\mathbf{W}_K$ still has the same dimensions as $\mathbf{W}$, but its entries have changed values, since (3.10) amounts to replacing N with K in (3.9):

$$w_{ij}^K = \sum_{k=1}^{K} p_{ik} e_{kj} \tag{3.11}$$

PCA and Cumulative Analogy. To see how PCA naturally incorporates the cumulative analogy scheme, let us look at the system in action. Consider a source

concept i which is in reality highly related to target concept j, while the Wikispeedia method leaves $d(i, j)$ undefined. Now consider also a set $\mathfrak{C}$ of source concepts (i.e., a set of row vectors of $\mathbf{W}$) which are similar (in a vector sense) to $\mathbf{w}_i$. These source concepts will reside in a part of concept space close to $\mathbf{w}_i$, so they will project similarly onto eigenconcepts (because a rotation will preserve the neighborhood structure of these concepts). If many of the source concepts in $\mathfrak{C}$ are highly related to j, the eigenconcepts on which they cause a significant projection will also be highly related to j. These eigenconcepts will cause the value w_{ij}^K to increase, compared to w_{ij}, so after smoothing, i is highly related to j. The fact that this was not the case to begin with is, in this case, directly attributed by our method to noise, caused by projecting onto insignificant eigenconcepts.

The number K of eigenconcepts is an important parameter. The larger we choose it, the more the reconstruction resembles the original data matrix. An empirical study of the effect of the choice of K, alongside examples, is provided on page 43.

An important property of the PCA-based smoothing method, and the reason why we call it a *first-order method,* is that it preserves the 'semantics', or 'meaning', of the matrix. For instance, $\mathbf{W}_K$ still contains relatedness values that *could have* originated from Wikispeedia games and that are comparable with distance values computed on the basis of actual games. Also, the relatedness measure is still asymmetric. In the next section we will see a similarity measure which does not preserve the 'meaning' and asymmetry of the matrix.

3.4.3 Second-Order Methods

We have just seen how PCA, implicitly, represents a source concept as the vector of its relatedness values to the ensemble of target concepts, and when giving an intuition for the PCA approach, we already referred to a notion of similarity between

such vectors. So one might as well interpret this vector similarity as an alternative measure of semantic relatedness. We will refer to this as a *second-order method* because, instead of considering relatedness directly as defined by the Wikispeedia measure, such an approach defines two concepts as related if they are related to similar sets of target concepts according to a first-order method.

When comparing two concepts this way, we are not considering the relatedness of a source concept with a target concept any more. Rather, we are measuring how similar the two concepts are when they are both playing the role of source concepts. In this sense, the 'meaning' of a second-order method is different from that of a first-order method. Also, since vector similarity is symmetric, this paradigm induces a symmetric relatedness measure. This is different from the first-order measures $\mathbf{W}$ and $\mathbf{W}_K$, which are both asymmetric.

Cosine Measure

We still have not discussed how exactly to quantify vector similarity. Standard in text mining and information retrieval is the so-called cosine similarity [Manning *et al.*, 2008], which measures the cosine of the angle between the two vectors to be compared. More formally, let $\mathbf{W}^{\cos}$ be the relatedness matrix based on the cosine measure, then

$$\mathbf{W}^{\cos} = \left(w_{ij}^{\cos}\right) = \left(\frac{\mathbf{w}_i \mathbf{w}_j^{\mathrm{T}}}{\|\mathbf{w}_i\| \, \|\mathbf{w}_j\|}\right), \tag{3.12}$$

where $\mathbf{w}_i$ is the i-th row of $\mathbf{W}$. All values are from $[-1,+1]$ (or from $[0,1]$ if we do not mean-center $\mathbf{W}$, since then all entries are non-negative).

Combining Dimensionality Reduction and Cosine Measure

The two ideas described above, PCA and second-order relatedness using the cosine measure, can be combined. Instead of representing source concepts as rows of $\mathbf{W}$ for the purpose of comparing them using the cosine measure, we may first project

the source concepts (i.e., row vectors) into reduced eigenspace and subsequently measure the cosine of the angle between rows of the eigenspace representation $\mathbf{P}_K$ (cf. (3.10)). Modifying (3.12), we get the new cosine relatedness matrix

$$\mathbf{P}_K^{\cos} = \left(\frac{\mathbf{p}_i^K (\mathbf{p}_j^K)^{\mathrm{T}}}{\|\mathbf{p}_i^K\| \, \|\mathbf{p}_j^K\|} \right), \tag{3.13}$$

where $\mathbf{p}_i^K$ is the i-th row of $\mathbf{P}_K$. This is similar to the approach also taken by Latent Semantic Analysis [Landauer and Dumais, 1997] (on a different kind of input matrix, of course).

Using $\mathbf{P}_K^{\cos}$ instead of $\mathbf{W}^{\cos}$ has the computational advantage that computing an entry (a dot product) requires only $\Theta(K)$ instead of $\Theta(N)$ operations.

Note that deploying the cosine measure on $\mathbf{P}_K$ is equivalent to doing so on $\mathbf{W}_K$, the result of the first-order method. This is because $\mathbf{w}_i^K = \mathbf{p}_i^K \mathbf{E}_K$ (cf. (3.10)), such that reformulating the numerator of (3.13) for $\mathbf{W}_K$ instead of $\mathbf{P}_K$ yields

$$\mathbf{w}_i^K (\mathbf{w}_j^K)^{\mathrm{T}} = \mathbf{p}_i^K \mathbf{E}_K \mathbf{E}_K^{\mathrm{T}} (\mathbf{p}_j^K)^{\mathrm{T}} = \mathbf{p}_i^K (\mathbf{p}_j^K)^{\mathrm{T}},$$

since $\mathbf{E}_K \mathbf{E}_K^{\mathrm{T}}$ is the identity matrix (because the eigenconcepts are pairwise orthonormal). For the denominator, we have the equality

$$\|\mathbf{w}_i^K\| = \sqrt{\mathbf{w}_i^K (\mathbf{w}_i^K)^{\mathrm{T}}} = \sqrt{\mathbf{p}_i^K (\mathbf{p}_i^K)^{\mathrm{T}}} = \|\mathbf{p}_i^K\|,$$

and analogously $\|\mathbf{w}_j^K\| = \|\mathbf{p}_j^K\|$.

3.4.4 Results

Having delineated both first- and second-order techniques for increasing the coverage of our relatedness measure, we will now evaluate their performance. The goal of these experiments is to determine which method—first-order PCA smoothing, second-order cosine measure without PCA, or second-order cosine with PCA—correlates best with the human notion of semantic relatedness, as captured in a

ground-truth test set, and which eigenspace dimensionality K yields the best results. We also deliver one concrete example to showcase the effect of PCA smoothing on the Wikispeedia relatedness matrix. Finally, we investigate the effect of the number of games in the data set on the quality of our relatedness measure.

For this evaluation, Wikispeedia relatedness, as expressed in $\mathbf{W}$, was computed based on 19,000 game traces. The game was played on the schools edition of Wikipedia described in Section 2.3.

Performance and Optimal Eigenspace Dimensionality

As ground truth, we use the human-defined WordSimilarity-353 test collection [Finkelstein *et al.*, 2002; Gabrilovich, 2002]. This data set contains 353 word pairs alongside relatedness[3] values (between 0 and 10) attributed to them by a number of human raters (for some pairs 13, for others 16). We take the mean of all human labels for a pair as the ground-truth relatedness of that pair and measure performance as the correlation coefficient between the ground truth and the relatedness values computed by our methods.

A useful beacon is the average pairwise inter-human correlation. It is only 0.61, which shows that semantic relatedness is a notion on which there is no overwhelming consensus among humans. It is important to bear this in mind when judging the correlation with humans as achieved by a computational method.

Of the 353 pairs we can use as a test set only those whose constituent concepts also have articles in the small Wikipedia version used in our proof of concept.

[3] Its name notwithstanding, WordSimilarity-353 does not measure similarity but relatedness, in the terminology of Section 3.1. For instance, test takers are asked in the instructions to attribute a high value to an antonymous concept pair.

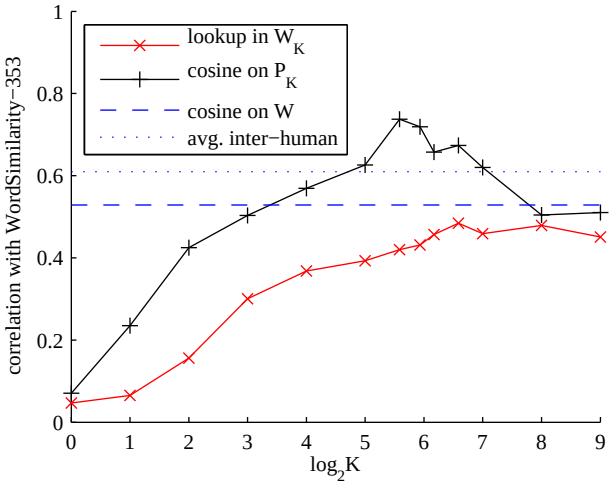


Figure 3–3: Performance of the first- and second-order methods for increasing coverage as functions of eigenspace dimensionality K (log scale).

There are 39 such pairs; they consist of 59 distinct concepts. All of these concepts appeared in Wikispeedia games, but only twelve pairs have a distance defined by the original Wikispeedia method of Section 3.3.1. The purpose of the methods discussed here is exactly to deal with this sparsity.

Figure 3–3 plots performance as a function of dimensionality, with K ranging from 1 to 512. Over the test set of 39 pairs, $\mathbf{W}^{\cos}$, the cosine measure applied to $\mathbf{W}$, achieves a correlation of 0.53 with the ground truth. This is significantly improved upon by using $\mathbf{P}_K^{\cos}$, the cosine similarity on $\mathbf{P}_K$ instead. At $K = 48$ we reach the optimal correlation of 0.74. As K tends to N, the performance of $\mathbf{P}_K^{\cos}$ approaches that of $\mathbf{W}^{\cos}$, since $\mathbf{W}_N = \mathbf{W}$ and applying the cosine to $\mathbf{P}_K$ is equivalent to applying it to $\mathbf{W}_K$, as explained above.

The fact that we surpass the average inter-human correlation means that our method agrees with the average human better than two humans agree with one another on average.

The first-order PCA smoothing method attains a performance slightly worse than the cosine measure on $\mathbf{W}$, correlation being 0.48 at the optimal $K = 96$. Note

Without PCA (i.e., $K = N = 4{,}604$):	
Board game	0.9921
Card game	0.9779
Chess	0.9729
Blackjack	0.9660
Pac-Man	0.9563
Doctor Who	0.9460
Playing card	0.9275
Computer and video games	0.9256
Commodore 64	0.9145
James Bond	0.8852
Nintendo	0.8820
Japan	0.8091

$K = 512$:	
Computer and video games	1.1094
Nintendo	0.9625
Chess	0.9183
Japan	0.8114
Doctor Who	0.7330
Board game	0.6975
Pac-Man	0.6108
Card game	0.5781
Blackjack	0.5364
Playing card	0.5156
James Bond	0.4704
Commodore 64	0.3607
Dice	0.2502
The Lord of the Rings	0.1960
Comics	0.1808
Monopoly (game)	0.1711
Alchemy	0.1557
Norse mythology	0.1515
Douglas Adams	0.1400
Fiction	0.1285
Star Wars	0.1276
Birmingham	0.1247
BBC	0.1129
Go (board game)	0.1053
Floppy disk	0.0927
The Simpsons	0.0875
17th century	0.0828
Confucianism	0.0806
Windows Vista	0.0758
Advertising	0.0745

$K = 96$:	
Japan	0.8614
Computer and video games	0.4358
Television	0.3587
Nintendo	0.3013
Computer	0.2231
Electronics	0.1987
Nintendo Entertainment System	0.1871
Sony	0.1558
Chess	0.1534
Super Mario Bros.	0.1330
Toy	0.1292
The Lion King	0.1169
Film	0.1115
Board game	0.1104
Animation	0.1087
Technology	0.1063
Mario	0.1041
Pac-Man	0.0949
The Simpsons	0.0901
Economics	0.0881
Game theory	0.0841
Planet	0.0832
Attack on Pearl Harbor	0.0791
Culture	0.0782
Sun	0.0697
Moon	0.0692
Education	0.0673
Automobile	0.0639
Ancient Greece	0.0591
Video	0.0586

Table 3–3: The 30 source concepts most closely related to the target concept GAME, according to $\mathbf{W}_K$, for different choices of K. Concepts printed in bold are those for which relatedness to GAME is defined by the Wikispeedia method.

that, since WordSimilarity-353 defines a symmetric relatedness measure, we, too, symmetrized the first-order method for the purpose of this evaluation, defining the relatedness between concepts i and j as $\max\{w_{ij}^K, w_{ji}^K\}$.

Effect of Dimensionality Reduction

To get a feel for the effect of PCA smoothing, let us analyze an example. We take the target concept GAME and look at the source concepts most related to it. The plain Wikispeedia method, without PCA smoothing, defines relatedness values for only twelve source concepts; they are shown on the left of Table 3–3, in order of decreasing relatedness.

Doing no PCA is equivalent to keeping all $N = 4{,}604$ eigenconcepts. The list in the center of Table 3–3 shows the result of keeping only the 512 most important eigenconcepts. There are three points to note: First, the twelve source concepts for which relatedness was originally defined stay on top of the list. Second, their order is scrambled. Third, many new concepts get their relatedness raised to a value

greater than zero; the highest-ranked amongst these is DICE, which is, as many others, certainly justified.

The change is more drastic when we discard even more eigenconcepts, keeping only the top 96, the optimal K according to the above evaluation. Now the original order is further confounded, to the extent that numerous new concepts mingle with the original ones in the upper part of the list and many of the original twelve even get pushed out of the top 30. Foregoing a formal evaluation involving human respondents, we state that the bigger portion of new concepts (e.g., SUPER MARIO BROS., TOY, GAME THEORY) are justified.

This example shows that, as expected, generalization is more aggressive the more we reduce dimensionality.

Effect of Number of Games

The previous results have all been based on a fixed data set consisting of 19,000 games. It is, however, also interesting to analyze how data set size impacts performance. Intuitively, a larger data set should make the relatedness measure more accurate. Figure 3–4 demonstrates that this is in fact the case, with the quality of the second-order measure steadily increasing as a function of the number of games, until it hits the correlation of 0.74 reported above.

The optimal K ranges between 30 and 60, which means it is relatively stable, given that theoretically all dimensionalities between 1 and $N = 4{,}604$ are possible (or slightly less if $\mathbf{W}$ does not have full rank).

3.5 Adjacency Matrix–based Methods

Nothing about the generalization methods just presented is specifically tailored to the Wikispeedia relatedness matrix $\mathbf{W}$. On the contrary, they may in principle be applied to any matrix of pairwise relatedness values. The goal of this section is

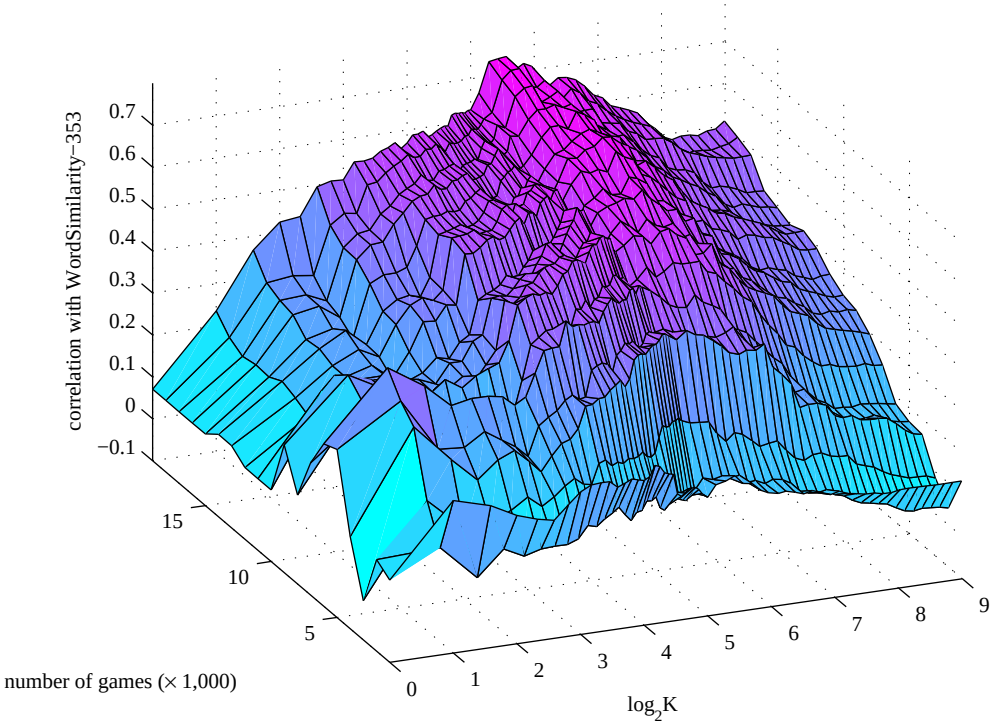


Figure 3–4: Performance of the second-order method $\mathbf{P}_K^{\cos}$ as a function of the eigenspace dimensionality K (log scale) and the size of the data set of Wikispeedia games.

to support this claim empirically. Concretely, we explore the use of Wikipedia's adjacency matrix as input to the second-order measure. Later on, Chapter 4 will show how the first-order method can be deployed in order to complete Wikipedia's hyperlink structure.

3.5.1 Wikipedia's Adjacency Matrix

The hyperlink structure of Wikipedia is captured completely in its adjacency matrix. Let N be the number of articles again. Then the adjacency matrix has N rows and N columns. The entry at position (i, j) is 1 if article i has a link to article j and 0 otherwise.

Recall from Chapter 1 that Wikipedia's hyperlinks may be considered noisy semantic links, a fact that is at the heart of the Wikispeedia game. Therefore, the adjacency matrix carries semantic relatedness information and can be deployed instead of $\mathbf{W}$ in all approaches of Section 3.4.

In this work we modify the adjacency matrix by weighting columns according to how many articles link to the respective article. This is useful because links pointing to an article that is rarely linked are more informative than links to articles that are linked from nearly everywhere else. For instance, in full Wikipedia, around 320,000 articles link to UNITED STATES OF AMERICA, while only 500 link to FEDERATED STATES OF MICRONESIA. The fact that an article links to FEDERATED STATES OF MICRONESIA is much more characteristic than that it links to UNITED STATES OF AMERICA. Let $\mathbf{A}$ be the weighted adjacency matrix. Its value at position (i, j) is

$$a_{ij} = \begin{cases} -\log(\text{indeg}(j)/N) & \text{if article } i \text{ links to article } j, \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

where $\text{indeg}(j)$ is the number of articles containing a link to j. Therefore, the term $-\log(\text{indeg}(j)/N)$ is the information content of the event 'picking an article that links to j' when we draw a Wikipedia article uniformly at random.

As we assumed for $\mathbf{W}$ before (cf. Section 3.4.1), we now assume that $\mathbf{A}$ has been mean-centered.

3.5.2 Cosine Measure on the Adjacency Matrix

The plain cosine measure without PCA (page 41) has previously been run on Wikipedia's adjacency matrix to compute semantic relatedness, by Ollivier and Senellart [2007] and Milne and Witten [2008a].[4] Both papers also modify the adjacency matrix using the information content weighting scheme introduced above. We now extend this approach by preprocessing the input matrix by means of PCA, just as we did for the case of the Wikispeedia relatedness matrix.

We test our method on two versions of Wikipedia:

1. The condensed Wikipedia Selection for schools which is also used in the Wikispeedia proof of concept (cf. Section 2.3).
2. A snapshot of full Wikipedia dating from March 6, 2009 [Wikipedia, 2009]; we used the Java toolkit WikipediaMiner [Milne, 2009], which maintains a database in the background to facilitate quick look-up of basic information such as the set of links contained in an article or pointing to it.

The implementation for the former is straightforward, while the latter offers some additional challenges: First, whereas we assumed a one-to-one mapping between words and concepts (i.e., Wikipedia articles) for the small version, disambiguation becomes necessary in the full version, since for most words there are many candidate articles (or *senses*) they could refer to; e.g., the phrase 'Monk' will refer most of the time to a male nun and correspond to the article MONK, whereas in a jazz-related context, it probably means THELONIOUS MONK. Second, PCA is intractable on the very large adjacency matrix of full Wikipedia. We now address these two issues before we present the results attained for both Wikipedia versions.

[4] It is referred to as 'cosine with tf-idf weight' by Ollivier and Senellart, and as 'TF×IDF inspired' by Milne and Witten.

Disambiguation in Full Wikipedia

To allow for a fair comparison, we use the disambiguation method from Milne and Witten's [2008a] 'final relatedness measure': First, we consider as candidate senses of a word (or sequence of words) all articles to which it ever links, i.e., for which it is an *anchor*. From this set, we purge all senses that receive less than 1% of the anchor's links. Now we list all pairs of candidates in order of decreasing relatedness (using the cosine measure that is the topic of this section) and keep only the pairs that are within 40% of the most related pair's value. Out of the remaining pairs, we return the one whose constituent articles receive the highest percentages of the respective anchor's links (the two separate percentages are summed to obtain a single value).

Finally, to also attribute high relatedness to words that often co-occur in a fixed phrase (e.g., 'bike' and 'path' in 'bike path'), the cosine relatedness value is increased by an additive term that captures the frequency with which the concatenation occurs as a link anchor in Wikipedia.

Making PCA Tractable on Full Wikipedia

The full Wikipedia dump contains over six million pages, 2,697,268 of which are actual articles (the rest are, among others, category, redirect, or disambiguation pages), and the database created by WikipediaMiner is 20 GB in size. Now $N = 2{,}697{,}268$ and consequently the $N \times N$ adjacency matrix $\mathbf{A}$ would occupy 29 terabytes of memory (assuming 32-bit floating point precision); a sparse representation is useless as well, because mean-centering turns most zeros of the sparse original adjacency matrix into negative numbers. So, in order to make PCA and thus our method tractable on full Wikipedia, we have to carefully shrink the adjacency matrix beforehands.

First, we reduce the size of the weighted, non-mean-centered adjacency matrix. In terms of columns, we keep only those associated with articles that have at least 15 incoming and 15 outgoing links. This way we eliminate articles about the most obscure topics—seemingly a majority of Wikipedia—, reducing the width of the data matrix to $w = 468{,}510$ (17% of the original width). The same method of constraining the set of articles is used by Gabrilovich and Markovitch [2007]. Remember that columns are the features of the data matrix, so discarding 83% of the columns could be described as feature selection.

To compress the height of the matrix, we keep a row only if the article it represents is about a topic for which the schools selection (version of 2008/9 [Wikipedia, 2008]) contains an article as well. This reduces the height of the data matrix to $h = 5{,}503$ (0.02% of the original height). Recall that rows are the data points of the data matrix, so discarding 99.8% of the rows amounts to shrinking the set of training samples for our algorithm aggressively, to only the most important articles (as determined by this other source of information). We will see in Section 4.5.1, when we use the first-order method for the task of link prediction, that restricting the set of training concepts that drastically does not impede performance on a set of test concepts that were excluded from the training process.

After decreasing the size of the matrix, we mean-center it and obtain the $h \times w$ matrix $\hat{\mathbf{A}}$. This matrix has a lot more columns than rows, which makes it amenable to a trick used in a seminal image processing paper on 'eigenfaces' [Turk and Pentland, 1991]. As mentioned in Section 3.4.2, the eigenconcepts are the eigenvectors of the data covariance matrix, which can be written as $\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}$. By definition, this means

$$\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{e}_k = \lambda_k \mathbf{e}_k, \tag{3.15}$$

for an eigenconcept $\mathbf{e}_k$ with associated eigenvalue λ_k.

Now consider the eigenvectors of another matrix, $\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}}$. Eigenvector $\mathbf{v}_k$ fulfills

$$\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k = \mu_k \mathbf{v}_k, \tag{3.16}$$

for eigenvalue μ_k. Left-multiplying by $\hat{\mathbf{A}}^{\mathrm{T}}$ yields

$$\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}(\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k) = \mu_k(\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k), \tag{3.17}$$

so each $\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k$ is an eigenvector of $\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}$. More precisely

$$\mathbf{e}_k = \hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k, \quad \lambda_k = \mu_k. \tag{3.18}$$

The crucial observation is that $\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}}$ is $h \times h$, i.e., 5,503 $\times$ 5,503 in our case, which means it fits into memory, making it possible to compute the eigenvectors $\mathbf{v}_k$ efficiently. Subsequently, we can find eigenconcept $\mathbf{e}_k$ simply as $\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{v}_k$.

We reiterate that, while the eigenarticles are computed based on a small set of only 5,503 'training articles', there is nothing that keeps our algorithm from being applicable to any novel input article not appearing in the training set. We demonstrate this generalization capability numerically in Section 4.5.1.

3.5.3 Results

Next we present the results of our evaluation of the adjacency matrix–based methods, for both the small and the full Wikipedia version.

Wikipedia Selection for Schools

The quality of the cosine measure on the 4,604 $\times$ 4,604 adjacency matrix $\mathbf{A}$ of the Wikipedia Selection for schools is shown in Figure 3–5 as a function of the eigenspace dimensionality K. The test setup was the same as for the experiments with the Wikispeedia relatedness matrix $\mathbf{W}$ (cf. Section 3.4.4).

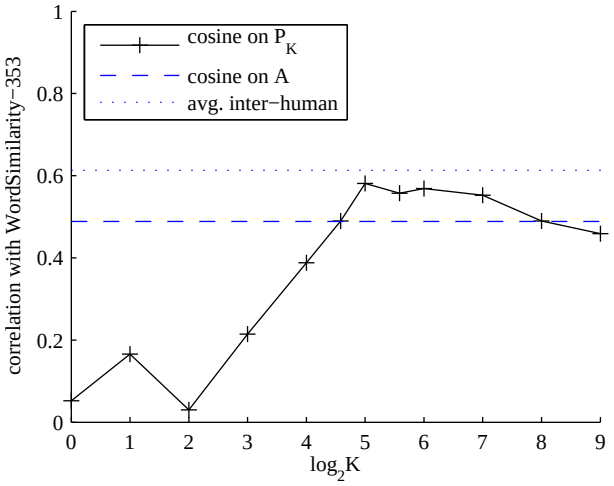


Figure 3–5: Performance of the second-order method when run on the adjacency matrix **A** of the Wikipedia Selection for schools, as a function of eigenspace dimensionality K (log scale).

Overall, the shape of the curve is similar to the one obtained by supplying **W** instead of **A** (cf. Figure 3–3). The main difference is that here the correlation with human ground truth peaks at 0.58, which is considerably lower than the 0.74 achieved with **W**. Without PCA, the plain cosine measure on **A**, too, performs worse, correlation with humans being 0.49 (as opposed to 0.53 with **W**).

We attribute this inferior quality to the fact that, as mentioned, relatedness as captured by the adjacency matrix is rather noisy. Wikispeedia was conceived as a tool to filter this noise, and we interpret it as justification for our approach that using the Wikispeedia relatedness matrix **W** yields more accurate measures of semantic relatedness than using the Wikipedia adjacency matrix **A**.

Optimal performance is reached for the dimensionality of $K = 32$. This is relatively close to the optimal value of $K = 48$ in the case of **W**, considering that the theoretically possible range is between 1 and $N = 4{,}604$ (or slightly less if **A** is not orthogonal).

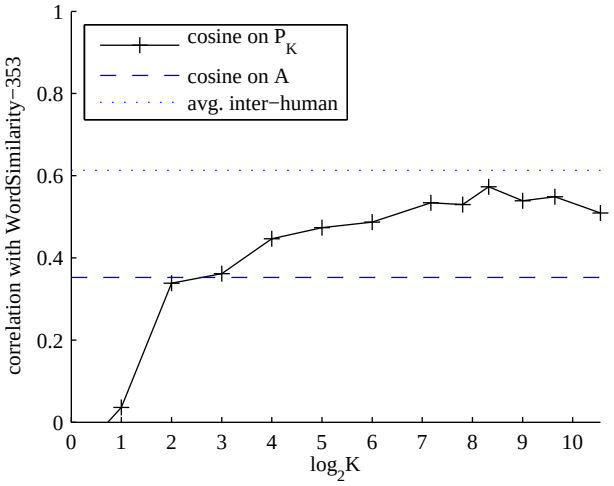


Figure 3–6: Performance of the second-order method when run on the adjacency matrix **A** of full Wikipedia, as a function of eigenspace dimensionality K (log scale).

Full Wikipedia

Figure 3–6 plots the result of the experiment when the very large adjacency matrix of full Wikipedia is used. Full Wikipedia covers many more topics than the schools edition, such that, out of the 353 concept pairs of the WordSimilarity-353 test set, 350 consist of concepts corresponding to article titles in full Wikipedia. Therefore, we may now extend our evaluation and base it on 350 concept pairs, rather than on the subset of 39 pairs we used before.

The plain cosine on **A** now yields a correlation of only 0.35, as opposed to the 0.49 achieved with the smaller matrix. This is probably due to increased noise in semantic terms: the number of articles in full Wikipedia is three orders of magnitude larger than in the schools edition, and since many of these are about minor topics, there is also a multitude of semantically meaningless links.

Still, PCA seems to recover from this noise, as the maximum performance is as high as in the case of the smaller matrix, at 0.57. Consequently, coverage, i.e., the number of concepts that can be compared, is much larger than for the schools edition, with virtually no loss in performance.

Not surprisingly, the optimal K is much greater for full Wikipedia, at a value of 320, since the inherent dimensionality of the large adjacency matrix is much higher than that of the small one. It is interesting to observe that the optimal dimensionality reported for LSA is nearly identical, at a value of 300 [Landauer and Dumais, 1997]. If this is more than a mere coincidence, it might suggest that this is the approximate dimensionality of human semantic space.

Milne and Witten [2008a] augment the plain cosine measure, which is based on out-links, by averaging it with an in-link–based measure inspired by the Normalized Google Distance [Cilibrasi and Vitányi, 2007]. In our case, this boosts the performance of the plain cosine measure from 0.35 to 0.57,[5] and that of the PCA-preprocessed cosine measure from 0.57 (as in Figure 3–6) to 0.66, for the optimal $K = 320$. The reason for this is the high quality of the in-link–based measure alone, which has a correlation of 0.64 with humans.

3.6 Discussion

To conclude this chapter, we now summarize the different techniques for inferring semantic relatedness we propose, discussing their respective strengths and disadvantages. We also put our approach in the context of previous work and finally discuss limitations and directions for future research.

3.6.1 Summary of Proposed Methods

In this chapter we present a novel method for computing the semantic distance (or its complement, relatedness) between concepts, based on data from the online game Wikispeedia, which builds on Wikipedia's hyperlink structure. This approach is

[5] Note that this is not the 0.57 shown in Figure 3–6; the identical numbers are but a coincidence.

inexpensive but cognitively plausible by directly extracting human common sense. Our measure is computed incrementally, is asymmetric, accounts for higher-order relatedness, and does not fulfill the triangle inequality, all of which are desirable from a cognitive viewpoint. It also has an information-theoretic interpretation.

Without any generalization, i.e., by just looking up values in the Wikispeedia relatedness matrix $\mathbf{W}$, the resulting measure is very precise and is consequently good for nearest-neighbor finding, as shown in Section 3.3.3. However, due to the sparseness of $\mathbf{W}$, it has low coverage and will therefore not be very useful when the task is to determine the semantic relatedness of two arbitrary concepts.

This original measure is equivalent to the first-order approximation when $K = N$. It can be spiced up with generalization *ad libidinem* by decreasing K. The generalized measure is still asymmetric, by maintaining the conceptual distinction between source and target concepts.

Employing a second-order method based on the cosine measure results in higher correlation with human test data. However, going from a first- to a second-order method changes the similarity measure drastically. The latter has no correlation with (i.e., no linear dependence on) the former,[6] is symmetric, and no more relates a source concept to a target concept but rather defines two source concepts as related if they are related to similar sets of target concepts according to the original Wikispeedia measure (when cosine is computed on $\mathbf{W}$) or the first-order measure (when cosine is computed on $\mathbf{P}_K$).

[6] We sampled ten sets of 1,000 random concept pairs each and compared, in each set, the relatedness values from $\mathbf{W}_{96}$ to those from $\mathbf{P}_{48}^{\cos}$. The average correlation across the ten sets is very close to zero at 0.015. The values of 96 and 48 for K are the optimal values found in the above evaluation.

The second-order measure comes closer to the notion of similarity (a specific type of relatedness) than the original Wikispeedia measure or the first-order generalization, in the nomenclature of Section 3.1: There, we cited Tversky [1977], who defines the similarity of two concepts as the overlap of their respective feature vectors. If we consider a concept's Wikispeedia relatedness values to all the other concepts (or eigenconcepts) to be its features,[7] then measuring vector similarity of rows in $\mathbf{W}$ (or $\mathbf{P}_K$) using the cosine metric amounts to computing the similarity of feature vectors, and second-order relatedness becomes a similarity measure. Intuitively, two concepts are then similar to each other if they are often triggered by the same third concepts in a person's mind (cf. discussion on page 26).

Second-order methods are better suited if the task is to rate the relatedness of two arbitrary given concepts (as opposed to finding the nearest neighbors of only one given concept), but if at the same time one wants to preserve approximately the structure of the original relatedness measure, one should stick to the first-order method using $\mathbf{W}_K$.

Since dimensionality reduction boosts the performance of the second-order cosine method, it is useful regardless of whether we use a first- or a second-order method.

To demonstrate that our generalization methods are applicable beyond the context of Wikispeedia relatedness, we also tested them on the Wikipedia adjacency matrix. The latter is amenable to our techniques because Wikipedia hyperlinks implicitly contain semantic relatedness information. However, as the semantic value

[7] Tversky [1977] explicitly includes the relation to other concepts when listing potential elements of a concept's feature vector: 'It includes appearance, function, relation to other objects, and any other property of the object that can be deduced from our general knowledge of the world.'

of Wikipedia links is rather noisy, our methods perform worse than when run on the Wikispeedia relatedness matrix **W**.

In the case of the second-order method, the optimal eigenspace dimensionality is 32 for **W** and 48 for the adjacency matrix of a small Wikipedia version. When the method is applied to the adjacency matrix of full Wikipedia, it is 320, i.e., an order of magnitude greater. Still, the method performs as well on full Wikipedia as it does on the small edition, despite the fact that the much larger coverage necessitates context-dependent disambiguation.

3.6.2 Relation to Previous Work

We will now place the methods we propose in the context of previous work. Most of the papers we refer to have already been mentioned in Section 3.2, in which we summarize related work.

Wikispeedia Distance. Ollivier and Senellart's [2007] method is based on the notion of random walks on a Markov chain. In our definition of Wikispeedia distance (cf. (3.3)), we, too, leverage random walks, by using PageRank as a normalization factor.

In Section 3.2 we have seen that several recent methods [Milne and Witten, 2008a; Veksler *et al.*, 2008] have drawn on Cilibrasi and Vitányi's [2007] Normalized Google Distance (NGD), which is an approximation of the uncomputable Normalized Information Distance (NID). Normally, NID is symmetrized, but the asymmetric definition would be $K(y|x)/K(y)$, where $K(y)$ is the Kolmogorov complexity of y (i.e., the length of the shortest program to output y) and $K(y|x)$ the conditional Kolmogorov complexity of y given x (i.e., the length of the shortest program to transform x into y) [Li and Vitányi, 2008]. Intuitively, this fraction is the percentage of y's information not yet contained in x. Kolmogorov complexity is uncomputable, but it can be approximated. NGD makes use of the number of

Google hits for the queries 'y' and 'x, y' for this purpose; Milne and Witten [2008a] leverage in-link statistics in Wikipedia. The distance of (3.3) can be understood as an approximation of the asymmetric NID, too: the numerator is the number of bits needed to encode a path from a_i to g, or in other words, to transform a_i into g (approximating $K(g|a_i)$), while the denominator is the *a priori* number of bits required to encode concept g (approximating $K(g)$).

In the Wikispeedia approach we exploit human click behavior in order to construct a measure of semantic relatedness. Kaur and Hornof [2005] expose an interesting symmetry: they invert the process and use existing measures of semantic relatedness to predict user click behavior on websites.

First-Order Method. Our first-order generalization method, although rather different in terms of problem domain, has been inspired by the AnalogySpace model of Speer *et al.* [2008] in terms of methodology. Whilst we are applying PCA to a sparse matrix of semantic relatedness values, they do so to a sparse matrix representing common-sense facts collected from humans (e.g., 'chopsticks are usually found in a kitchen'). The paradigm of cumulative analogy underlying both AnalogySpace and our first-order method can be traced back to Chklovski [2003].

Second-Order Method. As already mentioned in Section 3.5.2, the cosine measure on the adjacency matrix has been anticipated by Ollivier and Senellart [2007] and Milne and Witten [2008a], however, without the PCA preprocessing step.

Veksler *et al.*'s [2008] technique, although not based on Wikipedia, is also akin to our second-order method. Like them, we build a term–term matrix, but using the relatedness measure computed from Wikispeedia games instead of the Normalized Google Distance [Cilibrasi and Vitányi, 2007]. In the next step, however, we run PCA to reduce the dimensionality of the matrix, while Veksler *et al.* condense semantic space by selecting a subset of matrix columns using a genetic algorithm.

They also show that their vector-space model can be used for comparing entire documents. To represent a document as a vector, they sum the vectors representing the single words in that document; the relatedness between documents is then defined as the cosine similarity between their vectors. We did not conduct experiments with this technique but expect it to work equally with our second-order measure, given the analogous ways in which the respective matrices are constructed.

3.6.3 Limitations and Future Work

An interesting future experiment could test empirically whether indeed our method has the capability to measure the semantic relatedness between entire documents rather than merely between single concepts, which we anticipate because of its similarity to Veksler *et al.*'s approach, as just discussed.

In terms of limitations, it must be noted that, although we expect the Wikispeedia method for inferring semantic relatedness to also work on full Wikipedia, the proof of concept we evaluate here uses a small Wikipedia version. Instead of adapting our code to full Wikipedia, it would be reasonable to first conduct an exploratory analysis on data gathered by one of the alternative implementations of the Wiki Game (in case these data can be obtained), which all use full Wikipedia as a data set.

Wikispeedia hinges on the property that it tends to produce trajectories that have been guided by human common sense rather than shortest paths that optimally connect two articles in the Wikipedia link graph. However, there is currently no safeguard to prevent a malicious player from solving a given Wikispeedia mission using a shortest-path algorithm and following step by step the solution it returns. Although this worst case is far from being the average case—since actual solutions tend to be longer than optimal ones, as shown in Section 2.5.1—, it should ideally

be ruled out by the game design. In Section 2.5.2 we sketch a two-player version of Wikispeedia that could afford this.

A 'meta-parameter' in our first- and second-order methods is the algorithm used for dimensionality reduction. Throughout this thesis, we employ PCA for this purpose. A non-linear generalization of PCA using stacked autoencoders has been proven to outperform PCA on several tasks [Hinton and Salakhutdinov, 2006]. However, while eigenspace dimensionality is the only parameter to be chosen in PCA, the structure of such autoencoding architectures can be tuned in many more ways. Also, training procedures for weight learning still constitute an open research problem. For these reasons, we did not experiment with stacked autoencoders.

CHAPTER 4

Hyperlink Prediction through Dimensionality Reduction

In Section 3.5 we saw that Wikipedia's adjacency matrix can be employed as input to the second-order methods of Section 3.4. We shall now demonstrate that the first-order PCA method, too, is useful in the context of the adjacency matrix, for the purpose of finding missing hyperlinks in Wikipedia. Such an algorithm can help improve the user experience as well as make Wikipedia a better resource for artificial intelligence and data mining applications relying on its link structure.

The remainder of this chapter is structured as follows. Section 4.1 motivates the problem and gives an introductory example. In Section 4.2 we summarize previous related work. In Section 4.3 we succinctly formulate the concrete problem at hand, give an intuitive explanation of how and why the PCA-based first-order method works in our setting, and provide the algorithm. The experimental setup for evaluating it is described in Section 4.4, while Section 4.5 presents the results, showing that our approach outperforms the previous state of the art in a human user evaluation. Section 4.6 discusses our research in the context of previous work and points out limitations as well as avenues for future research. We conclude the chapter in Section 4.7.

4.1 Motivation

To maintain a consistent degree of quality, Wikipedia authors are encouraged to adhere to a *Manual of Style* [Wikipedia, 2010c; 2010b], which stipulates, among many other things, the following:

> ‘Provide links that aid navigation and understanding, but avoid cluttering the page with obvious, redundant and useless links. An article is said to be *underlinked* if subjects are not linked that are helpful to the understanding of the article or its context. However, *overlinking* is also something to be avoided, as it can make it harder for the reader to identify and follow those links which are likely to be of value.’ [Wikipedia, 2010b]

However, since humans are not flawless and the experience level varies widely among contributors, articles deviate frequently from these rules, which affects the textual content of articles as well as the hyperlinks they comprise. Consequently, human authors often forget to add links that should be there according to the editing guidelines. It would be desirable to detect such missing links automatically because it could enhance the browsing experience significantly. In the context of Wikispeedia, too, feedback from players has made it clear that frustration often results if a specific link is expected yet not to be found. In general, not only human readers but equally artificial intelligence and data mining programs that exploit Wikipedia’s link structure (such as those presented in this thesis) would profit from a data set that has been improved this way.

In this chapter we present an algorithm that has the capability of finding missing links in Wikipedia. As an example, consider the article about KARL MARX. It misses essential connections to other relevant articles, for instance it contains no links to SOVIET UNION or PROLETARIAN REVOLUTION. Our method is capable of predicting these links, as well as others. The top ten suggestions are listed in Table 4–1. (The link to SOCIALISM has actually been added to the online KARL MARX article since March 2009, the date of our local working copy of Wikipedia.)

We use the PCA-based first-order method of Section 3.4 in order to enrich existing articles with new links. Our approach can be viewed as using generalization

Suggested link target	Anchors	Gain
SOCIALISM	socialist, socialism	1032.9
SOVIET UNION	Soviet Union	939.7
DEMOCRACY	democratic	892.4
SOCIAL DEMOCRACY	Social-Democratic	826.2
JEW	Jewish, Jews	774.9
STATE	state, states	734.4
SLAVERY	slavery	726.3
POLITICS	political, politics	702.3
PROLETARIAN REVOLUTION	proletarian revolution	667.8
PROPERTY	property, private property	663.4

Table 4–1: Top ten suggestions for missing links to be added to the article about KARL MARX. Anchors are phrases on which the link can be placed. 'Gain' is the score for the suggestion (cf. Section 4.3).

from existing data in order to align articles to a more uniform linking policy. The intuition underlying our work is that of cumulative analogy (cf. page 39). Consider for instance Chuuk, Kosrae, Pohnpei, and Yap, the four states forming the Federated States of Micronesia. If most articles that link to CHUUK, KOSRAE, and POHNPEI also link to YAP, then another article that already links to CHUUK, KOSRAE, and POHNPEI but not to YAP should probably be modified by adding that missing link—provided the word 'Yap' occurs in the article.

4.2 Related Work

There have been several attempts to tackle the problem of suggesting links for Wikipedia.

Fissaha Adafre and de Rijke's [2005] approach can enrich articles that already contain some outgoing links and is based on the structure of the Wikipedia link graph. The method consists of two steps. First, it identifies a set of articles which are similar to the input article. Then, the outgoing links that are present in the similar articles but not in the input article are suggested to be added to the input

article. A link is only suggested if its anchor text in the similar article is also found in the input article.

In step one, similarity is defined in terms of incoming links. Intuitively, given two articles, if it is often the case that the same page refers to both articles, then the two articles will be considered similar. The actual implementation is more complicated, consisting of several steps harnessing the indexing feature of the custom search engine Lucene [Apache, 2009].

Another, more recent method was proposed by Mihalcea and Csomai [2007]. It differs from Fissaha Adafre and de Rijke [2005] and the work presented here in that its input is a piece of plain text (the raw content of a Wikipedia article or any other document). It operates in two stages: detection and disambiguation. First, the algorithm decides which phrases should be used as link anchors, then it finds the most appropriate target articles for the link candidates.

To detect link candidates, the best method they tried computes the *link probability* of candidate phrases and selects the top m of them, where m equals 6% of the number of words in the article (they determined the value of 6% empirically). The link probability of an n-gram T is defined as the number of Wikipedia articles containing T as a link anchor divided by the number of articles containing T. It is the prior probability of T being used as a link anchor given that it appears in an article. For instance, the n-gram 'big truck' has a link probability of 0%, whereas 'Internet' has link probability 20%, i.e., every fifth article that mentions the Internet contains a link to its article. In this approach, 'Internet' is likely to be linked again, while 'big truck' is considered to not be a useful link anchorage.

Once the anchors have been chosen, disambiguation is key, since many phrases have several potential meanings (cf. page 49). To decide the best sense of a phrase,

Mihalcea and Csomai extract local features from surrounding text and train a machine learning classifier from Wikipedia articles, which can serve as labeled examples since the links they contain are already disambiguated. The features are a set of words occurring frequently in the document, as well as the three words to the left of the candidate, the three words to its right, and their parts of speech. As output, the method attaches to each sense candidate a numerical value representing the confidence in this sense being the correct one.

A third method, proposed by Milne and Witten [2008b], consists of the same steps, but in swapped order. They first find the best sense of each phrase and only then decide which phrase to use as a link anchor.

To disambiguate a term, they look up the articles to which it points when it occurs as a link anchor in Wikipedia. They call the frequency of each potential target article (or sense) its 'commonness'. Then they find all the unambiguous terms in the document; these are the terms that link to the same target article regardless where they occur as anchors in Wikipedia. Then they compute the average semantic 'relatedness' between the candidate term and the unambiguous terms. While any relatedness measure could be plugged in, they use the one we delineated on page 55 [Milne and Witten, 2008a]. Finally, they train a machine learning classifier to combine commonness and relatedness and predict the most appropriate sense of each phrase.

After all phrases have been disambiguated, Milne and Witten decide which of them to use as link anchors, based on several features of the input article. These include, among others, the link probability of the candidate, its semantic relatedness to the context, how often it appears in the document, and in what positions. Again, all features are combined to train a machine learning classifier. This approach has better precision and recall than the predecessor by Mihalcea and Csomai, in the

task of predicting the hyperlinks of a Wikipedia article whose plain text is given as input.

While our technique and the approaches just summarized deal with very similar problems, the methodology we propose is rather different, building directly on the PCA approach proposed in Section 3.4.2.

4.3 Proposed Method

Concisely, the problem we are attacking here can be formulated as follows: The input consists of a Wikipedia article containing at least a few hyperlinks to other Wikipedia articles; the input article is represented as a vector of these outgoing links. The task, then, is to produce a list of Wikipedia articles to which the input article should also link (but does not yet); it is desirable that the output list be ranked according to meaningful numerical values representing the confidence in each suggestion.

We begin with the weighted, mean-centered Wikipedia adjacency matrix $\mathbf{A}$, as defined in Section 3.5.1. Recall that, before mean-centering, entry (i, j) of $\mathbf{A}$ is proportional to the information content of the link between articles i and j if such a link exists, and zero otherwise. We adopt the following nomenclature: Rows of $\mathbf{A}$ are referred to as articles (i.e., articles are represented entirely in terms of their outgoing links); the principal components of $\mathbf{A}$ are now called *eigenarticles* (we had named them eigenconcepts in the context of semantic relatedness); the original space before projecting into eigenspace is called *article space* (concept space in the context of semantic relatedness).

Our approach to link prediction uses articles as input to the first-order method as described in Section 3.4.2. That is, a row from $\mathbf{A}$ is first projected into reduced eigenspace and then back into article space, yielding the reconstruction $\mathbf{A}_K$. After the back-projection we compare an entry a_{ij} of $\mathbf{A}$ to its equivalent a_{ij}^K in $\mathbf{A}_K$. If

there was no link between articles i and j originally, but $a_{ij}^K \gg a_{ij}$, then our method predicts that the link should be added.

On page 39 we explained why this algorithm implements the cumulative analogy scheme. Briefly recapitulating, consider an article i which should link to article j but does not, and also a set $\mathfrak{A}$ of articles which are similar to article i, in terms of the other outgoing links. These articles will be located in a part of article space similar to i, so they will project similarly onto eigenarticles (because PCA merely performs a rotation). If many of the articles in $\mathfrak{A}$ contain j as an outlink, the eigenarticles on which they cause a significant projection will also link to j. This will cause the value a_{ij}^K to increase, compared to a_{ij}, so article j will be suggested as a link from i as well. Our method attributes its absence in the original article to noise, caused by projecting onto insignificant eigenarticles.

Note that no heuristic is involved in our method. It simply exploits the statistical properties of the set of already existing links. We emphasize again the particular flavor of the use of PCA here (as also, e.g., in Speer *et al.* [2008]). Typical PCA applications strive to minimize the reconstruction error while compressing the data through dimensionality reduction. In our paradigm, this 'error' is exactly what we exploit. To underline this, we should speak of *reconstruction gain* or generalization gain rather than reconstruction error.

Pseudocode for the method we just described is provided in Algorithm 1. The steps laid out above are followed directly. The article to be augmented is projected into reduced eigenspace, then back into article space. The output is a list of link suggestions, ordered by the reconstruction gain of the links, i.e., by how much more weight they have after the projections, versus before.

Of course, a link can be suggested only if the appropriate anchor term occurs in the text of the source article. In order to prune away nonsense terms and stopwords from the beginning, and thus speed up the algorithm, we consider as

Algorithm 1 Wikipedia link suggestion

Input: Article i, represented by its outlinks $\mathbf{a}_i$;
minimum link probability β
Output: Link suggestions for article i, in order of decreasing quality
Static: Eigenarticle matrix $\mathbf{E}_K$
$\mathbf{p}_i \leftarrow \mathbf{a}_i \mathbf{E}_K^{\mathrm{T}}$ (projection into reduced eigenspace)
$\mathbf{a}_i^K \leftarrow \mathbf{p}_i \mathbf{E}_K$ (projection back into article space)
$\mathbf{g}_i \leftarrow \mathbf{a}_i^K - \mathbf{a}_i$ (the reconstruction gain vector)
$\mathfrak{L} \leftarrow \emptyset$ (set of link candidates)
for n-grams T of text of article i **do**
  if T has link probability $> \beta$ **and** there is an article j about topic T
  and i has no link to j **then**
    Add j to $\mathfrak{L}$
  end if
end for
for $j \in \mathfrak{L}$, in order of descending g_{ij} **do**
  Suggest link from i to j
end for

potential anchors only n-grams whose link probability (cf. Section 4.2) is above a specified threshold β. As n-grams we choose all sequences of between one and four words. A value of $\beta = 6.5\%$ was empirically found to balance precision and recall optimally [Milne and Witten, 2008b], which is why we use this threshold in our implementation.

4.4 Experimental Setup

We ran our algorithm on two versions of Wikipedia. In this section we first describe these data sets and then proceed to detail the experimental procedures for evaluating our algorithm on them.

4.4.1 Data Sets

We used the following two data sets to evaluate our approach:

1. The March 6, 2009, data dump of the entire Wikipedia [2009] (cf. Section 3.5.2). Recall that it contains 2,697,268 articles, so $N = 2{,}697{,}268$ in this case.
2. The Wikipedia Selection for schools (cf. Section 2.3). We upgraded to the 2008/9 edition [Wikipedia, 2008] for these more recent experiments. It contains 5,503 articles (so $N = 5{,}503$) and redirects were resolved more rigorously than in the 2007 edition (e.g., links to MÜNCHEN were changed to MUNICH, since the two are different titles of the same article) [Cates, 2009].

While the Wikipedia Selection for schools serves well as a proof of concept and for evaluating the potential of the technique, the full version of Wikipedia is certainly more interesting, for several reasons. First, Wikipedia's live online version is consulted by many Internet users on a daily basis. So, if our method can improve full Wikipedia, it will have much more traction than if it works only on a small subset of articles. Second, live Wikipedia is evolving constantly, articles being added or modified constantly. Thus, if our method is applicable to full Wikipedia, then it can be used by authors every day to find links they have probably forgotten to include in the articles they are writing. Third, Wikipedia contains over two million articles (three orders of magnitude more than the school selection). In order to cope with such a challenging amount of information, our algorithm really has to scale well. Finally, previous methods use full Wikipedia as a data set, and we want to compare the performance of our technique directly to them.

The adjacency matrix of full Wikipedia is too large to be kept in memory. Therefore, to make PCA tractable, we preprocess the matrix the same way as described in detail in Section 3.5.2. Once the eigenarticles have been computed (we use eigenspace dimensionality $K = 1{,}000$ for the full Wikipedia data set), Algorithm 1 can be run. Our implementation uses Java and the WikipediaMiner toolkit [Milne, 2009] (cf. Section 3.5.2).

On the contrary, the Wikipedia Selection for schools is small enough such that the entire adjacency matrix fits into memory. Computing the eigenarticles and implementing Algorithm 1 is then straightforward using Matlab's built-in functions. For this data set we choose an eigenspace dimensionality of $K = 256$.

4.4.2 Evaluation Method

We employed these two data sets in two different types of evaluation: using full Wikipedia, we show that the top link suggestion of our method is of high quality, while we employ the schools edition to demonstrate that the reconstruction gain computed by our algorithm is indeed an indicator of the quality of a suggestion.

Full Wikipedia

In the case of full Wikipedia, we evaluate the quality of our highest-ranked link suggestion by querying human raters on Amazon Mechanical Turk [Amazon, 2009]. In each rating task we presented the human contributor with the text of a randomly selected Wikipedia article about a topic T. The article text still contained the original outgoing links. The task description read as follows:

> 'You are presented with the text of a Wikipedia article about T.
>
> Below the article text, you are given the titles of four other Wikipedia articles. The article about T could potentially contain a link to each of these four articles.
>
> Your task is to identify the one link (from the list of four) which you consider most useful. A useful link should lead to an article that is relevant for the article about T, and which readers of the article about T would likely want to investigate further.
>
> In case you are not familiar with T, please make sure you get an idea of who or what T is by looking through the article text.'

In order to be able to compare our algorithm to Milne and Witten's, the definition of a useful link is directly copied from their instructions to human raters [Milne and Witten, 2008b], which in turn capture Wikipedia's linking policy [Wikipedia, 2010c].

The four outgoing links between which raters had to choose were the following:

1. The top link suggestion S made by our method, using the $K = 1{,}000$ most significant eigenarticles. Note that the article always contained an appropriate anchor for suggestion S, and of course T itself was never chosen as a suggestion.
2. The top link suggestion S_{MW} made by Milne and Witten [Milne and Witten, 2008b], i.e., the one to which their system attributes the highest confidence value. Their code is included in the WikipediaMiner toolkit [Milne, 2009] and could thus be used off the shelf.
3. A pre-existing link S_{PRE} already present in article T, selected uniformly at random, but different from S and S_{MW}.
4. A link S_{RND} to an article that is not linked from T but that could potentially be linked because its title is one of the n-grams of T's plain text. Again, this is chosen randomly and different from S and S_{MW} (and from S_{PRE} by definition). This serves as a random baseline.

The order of the four choices was randomized, to prevent any bias.

We evaluated the performance on a set of 181 articles randomly picked from the set of articles not used in computing the eigenarticles, to avoid overfitting and test whether our algorithm generalizes well to unseen data; call this set the *test set*. We constrained our random selection to articles with at least 100 incoming and at least 100 outgoing links. The reasoning is similar to that behind our choice of the columns of $\hat{\mathbf{A}}$ (cf. page 51): we wanted to ensure that the articles were not about

very obscure topics, so human raters would not have to read the article text in depth to be able to make an informed decision.

To facilitate the performance analysis, we considered only articles on which our method and that of Milne and Witten did not agree. (Out of the 200 articles we initially tried, the methods agreed on 8%.)

Each task was completed by six different raters, so the number of votes we gathered is $6 \times 181 = 1{,}086$. As a safeguard against participants who might potentially have clicked randomly rather than made an informed decision, we implemented a voting scheme that counts a vote only if it agrees with at least two others on the same task, which resulted in a set of 660 effective votes.

Wikipedia Selection for Schools

Since we test our algorithm on the full version of Wikipedia, we do not evaluate the quality of link suggestions the same way on the small selection as well. Instead, we focus on a more qualitative analysis. In particular, we demonstrate that links with small reconstruction gain are less useful than those with a large gain. This is desirable, since it implies that the numerical values our method attaches to link suggestions can be used to rank them in a meaningful way. The evaluation using full Wikipedia does not highlight this property of our algorithm, since there we only evaluate the quality of the *top* suggestion.

Here we make the assumption that a link suggestion from article i to article j is more likely to be valuable if the word 'j' appears in article i than if it does not. We call such links 'acceptable'.

Next note that the first **for** loop of Algorithm 1 considers only link candidates that could potentially be accepted because an appropriate anchor appears in the text of the source article. This is exclusively for reasons of efficiency. Nothing prevents us from looping over *all* potential target articles (i.e., all Wikipedia articles),

regardless of whether there is an apt anchor or not. This way we can first collect all link predictions and calculate later for what percentage of them a fitting anchor exists. According to the above assumption, the higher this percentage, the better the average quality of the suggested links.

In particular, we proceed as follows. We first compute all link suggestions, across the entire Wikipedia Selection for schools, which is easily done since the complete adjacency matrix fits into memory. Then we list the suggestions in order of decreasing reconstruction gain. Finally, we descend in this list, considering at each step a window of 10,000 consecutive entries. For each window, we compute the percentage of acceptable links. This way we obtain a running average of link suggestion quality.

4.5 Results

Having clarified the experimental procedures and the rationale behind them, we now present the results of our evaluations.

4.5.1 Full Wikipedia

The results of our evaluation using the full Wikipedia data set are summarized in Figure 4–1 (all gathered data can be found online [West, 2009a]). Our method won most votes (36%), followed by Milne and Witten (27%), the random pre-existing links (25%), and finally the baseline of random n-grams (11%).

Thus, our method outperforms the previous state of the art. Our top suggestion is considered best 9% more often than theirs. A difference of at least 4% is statistically significant at the $p < 0.05$ level (estimated by bootstrap resampling).

Also, the fact that our suggestions won significantly more votes than the randomly picked pre-existing links (11% difference; at least 6% is significant at the

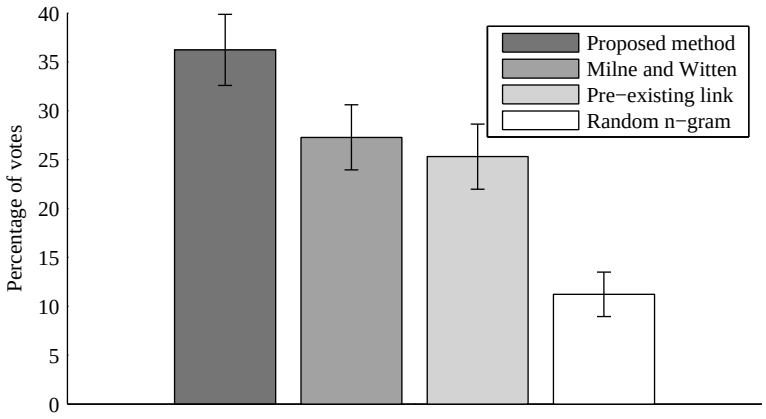


Figure 4–1: Results of the human user evaluation, in terms of percentages of votes won by the different link types (explained in Section 4.5.1). The error bars show the 95% confidence intervals (estimated by bootstrap resampling).

$p < 0.05$ level) implies that the top links our method finds are better than the average human-added link: we do not just find minor links that happen to have some relevance for the article being augmented; instead, we find important links that the human authors forgot to include.

The voting scheme we use to exclude spurious human raters is justified *a posteriori* by the low performance of the random baseline, which is according to our expectations.

This quality of suggestions is reached on a set of test articles that were not used in the eigenarticle calculation, which implies that our algorithm generalizes well to articles it was not trained from. This is crucial because it justifies selecting only a small subset of all Wikipedia articles as rows of $\hat{\mathbf{A}}$ (cf. page 51), a restriction without which PCA on the enormous adjacency matrix would be computationally infeasible.

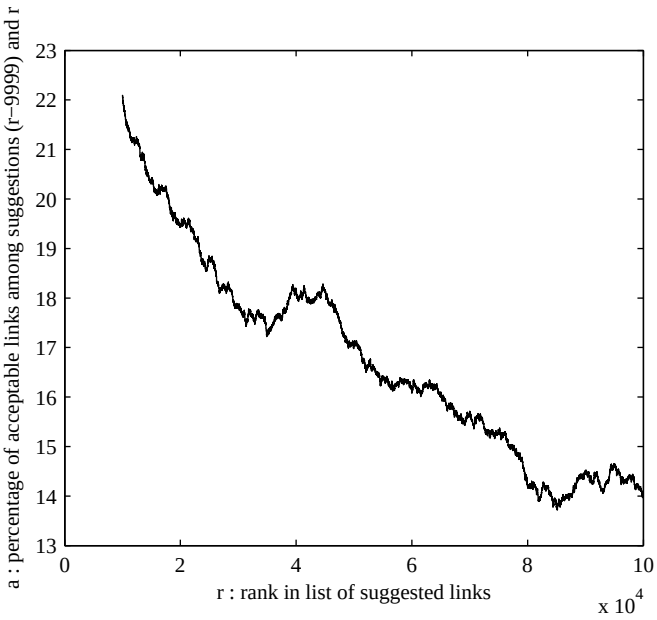


Figure 4–2: Running average of the number of suggestions that are acceptable because an appropriate anchor for the target appears in the source article. Note that the maximum is deceptively low because the running average is taken over 10,000 consecutive ranks.

4.5.2 Wikipedia Selection for Schools

Above we have described how we evaluate the validity of reconstruction gain as an indicator of link suggestion quality. Figure 4–2 plots average link quality as a function of rank in the list of link suggestions. The r-axis shows the rank; the a-axis shows the percentage of acceptable suggestions in the window of 10,000 suggestions up to rank r. The fact that this percentage decays as we descend in the list of suggestions means that fewer and fewer of the predicted links have an anchor in the source article, which in turn implies that the quality of suggestions decays as well, according to the assumption we made on page 73. We conclude that our algorithm does not just roughly separate good from bad suggestions but also ranks them continuously in a sensible way.

Note that the probability of a random article name appearing in the text of another random article is only 1.5% (estimated from 10,000 randomly selected article pairs), significantly lower than the 14% that Figure 4–2 shows for suggestions

90,001 to 100,000. This means that not only our *top* suggestions are much better than random ones (as shown in Section 4.5.1) but that this is true even far down in our ranking.

To illustrate the effect of our technique on more than a few hand-picked examples, we augmented a complete local copy of the 2008/9 Wikipedia Selection for schools by adding the 17,000 highest ranking links suggested by Algorithm 1 (an average of approximately three new links per article). The result can be browsed online [West, 2009a].

4.6 Discussion and Future Work

In this section we discuss our approach in the context of previous work, delineate ways in which the two could be combined, and point out avenues for future research.

4.6.1 Comparison to Previous Methods

To highlight the contributions of this research, we will now contrast it with the existing methods referenced in Section 4.2.

The technique coming closest to ours is that of Fissaha Adafre and de Rijke [2005], since it is based on the links rather than the text that articles contain. However, there are several important differences.

Fissaha Adafre and de Rijke gauge the similarity of two articles in terms of how many incoming links they share. To augment an input article with new links, they copy links from any *single* article that is sufficiently similar to the input article according to this measure. Our method represents articles in terms of their outgoing links and incorporates the cumulative analogy paradigm: 'If there are *many* articles sharing a lot of features (outlinks) among each other and with the input article, and if these articles also share a certain single feature (outlink), then the input article

should have that feature (outlink), too.' The fact that many similar articles, rather than just a single one, are required makes the method more robust to noise.

In addition to this robustness concerning where to copy from, our technique is also more careful regarding what to copy. If an article is similar enough to the input article, Fissaha Adafre and de Rijke copy any of its outgoing links, as long as the appropriate anchor text occurs in the input article. On the contrary, our method works with numerical values and can thus weight outlinks with importance values (reconstruction gain).

Also, the approach we propose naturally incorporates the two steps (picking the similar articles and ranking the candidate links before suggesting them for the input article) into one simple mathematical operation, PCA. Fissaha Adafre and de Rijke's first step alone seems considerably more complicated, involving a scheme of several rounds of querying the search engine that indexes the incoming links of each article.

Before we compare our method to Mihalcea and Csomai [2007] and Milne and Witten [2008b], we will first summarize their principal properties (for more details, see Section 4.2) in a concise list:

1. Both methods consist of **two separate phases**, link detection and link disambiguation.
2. They rely heavily on **several hand-picked features**, used to train machine learning classifiers.
3. These features strive to capture the **textual content** of the article to be augmented.

We demonstrated that we can outperform the state of the art [Milne and Witten, 2008b] with an algorithm that elegantly integrates detection and disambiguation in one **single phase**. To illustrate this, it is worthwhile to point out a subtlety we have glossed over in the pseudocode of Algorithm 1. We wrote 'topic T' in the first loop,

while in fact T is an n-gram, i.e., a sequence of words, which could be ambiguous. However, mapping the n-gram to the most appropriate target article is easy: given a source article i, the PCA will already have computed a score (the reconstruction gain) for *every* other Wikipedia article, so to retrieve the most appropriate sense of the n-gram T in article i, we simply look at all possible senses (all articles the anchor T ever links to in all of Wikipedia) and define 'topic T' as the one with highest reconstruction gain for source article i.

Even if the features used in the two approaches make sense intuitively and turn out to work well, they still had to be defined 'manually' by experts. On the contrary, our method is **featureless**. It merely completes the hyperlink structure of a document collection by means of a mathematically sound and proven generalization technique. There is no need to 'force' the algorithm to follow Wikipedia's linking policy [Wikipedia, 2010c] by hand-crafting features that encode those rules. Our technique starts from whatever linking policy is in place—most articles abide by it very closely to begin with—and enforces it where it is infringed, by eliminating the noise such a deviation represents.

The algorithm we propose works on the **hypertextual content** of an article (the set of outgoing links), not on its raw text. No advanced scanning or even parsing is necessary, as in the two text-based methods (e.g., Milne and Witten [2008b] need to know at what position in the article a phrase occurs; Mihalcea and Csomai [2007] even require part-of-speech tagging). We only ever inspect the article content in one trivial way, to see which n-grams it contains (and once, offline, to calculate link probabilities; but as explained in Section 4.3, this is not even integral to our approach but just a means of speeding up the algorithm). Our technique is based entirely on the link structure of the document collection. This rich source of information is not leveraged by Mihalcea and Csomai. Milne and Witten do use link structure, but more indirectly, to compute their semantic relatedness measure.

However, since it is used as a black box, this component could be replaced with any such measure and is not an integral ingredient of their approach.

4.6.2 Optimal Eigenspace Dimensionality

It should be mentioned that the memory requirements of our algorithm can be rather high, depending on how one chooses the eigenspace dimensionality K, since the eigenarticles have to be stored in RAM. Memory usage grows linearly in K.

Note that, while we have identified the optimal eigenspace dimensionality for the task of computing semantic relatedness, we have not fine-tuned this parameter in the case of Wikipedia link prediction. This is due to the fact that there is no reliable ground-truth test set for the latter task, so we had to resort to an *ad-hoc* human user evaluation, which prevented us from re-running the evaluation many times with varying parameter settings.

4.6.3 Synergies

Content-based methods have the advantage of being able to add links to raw text rather than documents that already come with a set of Wikipedia links. Our method does not have this capability. This is why it is important to point out that in the big picture our algorithm is not so much an alternative to Mihalcea and Csomai's [2007] and Milne and Witten's [2008b] as rather a tool to exploit dimensions of Wikipedia unaccounted for by those predecessors. Consequently, we conjecture that a combination of textual and hypertextual methods might have a synergetic effect: while in this paper we restricted ourselves to showing that our technique works well for suggesting links within Wikipedia, the method is applicable, without any changes, to any input document containing a basic set of links to Wikipedia. It could thus employ a text-based link suggester such as Mihalcea and Csomai's [2007] or Milne and Witten's [2008b] as a preprocessor and fill in links those methods have missed.

We conducted preliminary experiments with this approach but do not report results, for lack of a formal evaluation.

One could even go further and couple a textual technique with our hypertextual one, in order to link a complete plain-text document collection (such as a large news story archive) to Wikipedia in three steps: First, add a basic set of links to each document by means of a text-based technique. Second, compute the eigenarticles for this document collection. Third, run our method on all articles to complete the link structure. Step two only serves the purpose of fine-tuning the method to the characteristics of the document collection at hand. Alternatively, the eigenarticles computed from Wikipedia can be used.

A synergetic effect may also be expected when our method is deployed in a feedback loop: as Wikipedia authors accept (or reject) an increasing number of link suggestions, Wikipedia will comply ever closer to its own linking policy, which in turn means more accurate training data for the next generation of suggestions. A similar argument could be made for the text-based methods, yet it is more immediate for our approach, since it takes its own output—link structure—directly as input.

4.6.4 Detection of Missing Topics

While link suggestion is useful in its own right, the reach of our technique goes beyond. Recall that, unlike the existing approaches, our algorithm computes scores not only for phrases appearing in the input article but for *every* Wikipedia article. Let us take Table 4–2 as an example. It shows the top 15 suggestions of Algorithm 1 for the STATISTICS article of the Wikipedia Selection for schools.[1] Note that many links (those not marked with a star) could not be suggested for the sole reason that

[1] For reference, this article is reproduced in Appendix B.

	Suggested link target	Gain
	RANDOM VARIABLE	3.232
★	VARIANCE	2.819
	PROBABILITY DISTRIBUTION	2.469
	MEDIAN	1.800
	REAL NUMBER	1.454
	POISSON DISTRIBUTION	1.450
	EXPONENTIAL DISTRIBUTION	1.447
	BINOMIAL DISTRIBUTION	1.385
	CHI-SQUARE DISTRIBUTION	1.353
	PSYCHOLOGY	1.145
	PHYSICS	1.079
★	ENGINEERING	1.031
★	ECONOMICS	1.018
	COMPUTER SCIENCE	0.991
	ARITHMETIC MEAN	0.926

Table 4–2: Top 15 suggestions of Algorithm 1 for missing links to be added to the article about STATISTICS in the 2008/9 Wikipedia Selection for schools. 'Gain' refers to reconstruction gain. Links marked with a star could actually be added because the appropriate anchor text occurred in the source article.

there was no appropriate anchor text in the source article. It is interesting to see that, more often than not, it would be desirable if the article about STATISTICS did in fact cover the target topic. For instance, it is well possible that the author simply forgot to properly introduce the concepts RANDOM VARIABLE and PROBABILITY DISTRIBUTION or to mention that STATISTICS is of foremost importance to modern PHYSICS.

Consequently, our method can be deployed not only to suggest missing links but also to suggest missing topics. This feature, too, distinguishes our method significantly from previous link suggestion methods [Mihalcea and Csomai, 2007; Milne and Witten, 2008b]. They constrain their suggestions to topics that are present in the source article in the first place, and are thus unable to predict which topics *should* be present. They can only decide whether a term that already appears

in the article text should be used as a link anchor. Previous methods are discriminative topic detectors, ours is at heart a generative topic suggester.[2] In one potential application, our algorithm could be run on existing Wikipedia articles and point out those that could be improved by extending them to include specific additional topics.

4.6.5 Concept Clustering

The central computation of our algorithm is the projection of an article onto the eigenarticles. To understand the effect of this operation graphically, let us take a quick peek into eigenspace. Figure 4–3 plots 200 articles selected randomly from the full Wikipedia version, neglecting all higher dimensions and showing only the projections onto the two most important eigenarticles. In the notation of Section 3.4.2, the axes of the plot are $\mathbf{e}_1$ and $\mathbf{e}_2$, and article i has co-ordinates (p_{i1}, p_{i2}). The dashed line shows that the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$ is 'semantically separable': articles below the line are nearly exclusively about science-related topics, whereas those above the line live in the realm of the arts and humanities (history, culture, etc.).

This is a consequence of the fact that PCA finds the directions of largest variance in the data. Since a data point is defined by the outgoing links of an article and since articles about science topics typically have a very different set of outlinks from articles about the arts and humanities, these two classes are far apart in the subspace spanned by the first principal components of the data. These observations suggest that our method may also be used to cluster concepts into semantic classes.

[2] Although Fissaha Adafre and de Rijke [2005] do not mention it, we conjecture that their technique, too, is in principle able to suggest topics.

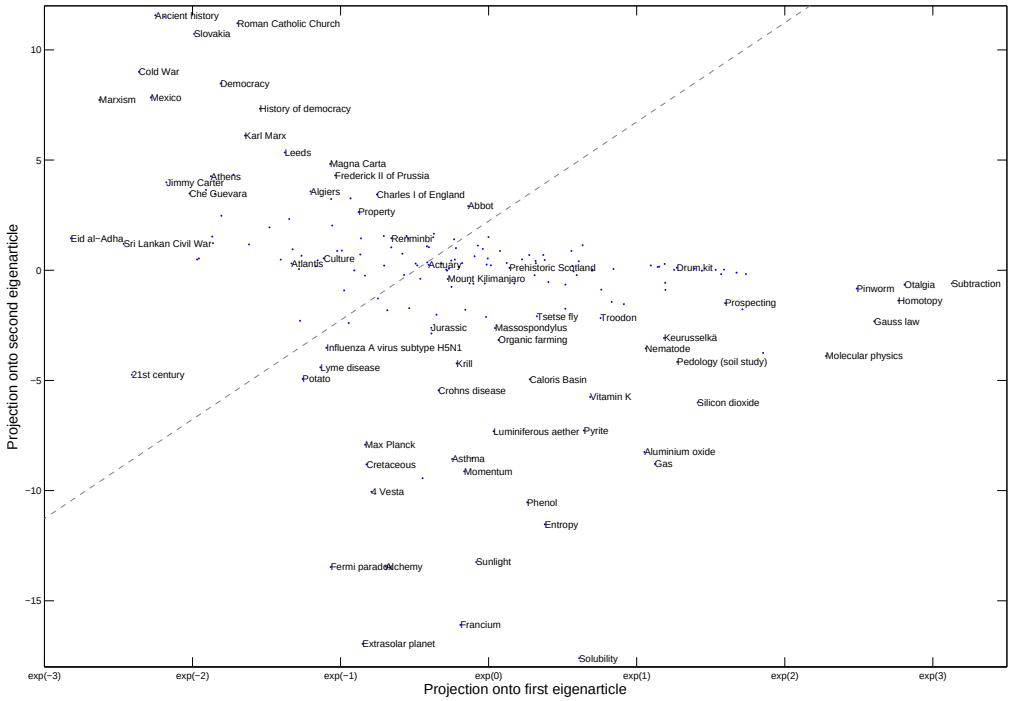


Figure 4–3: Projection of 200 randomly selected articles onto the two principal eigenarticles. To increase legibility, only a subset of points is labeled, and the x-axis is logarithmic (no log transformation could be performed on the y-axis, since the logarithm is not defined for the negative values). The dashed line roughly separates articles about the sciences from those about the arts and humanities.

4.6.6 Removing Links

Our algorithm measures the quality of a link suggestion in terms of its reconstruction gain. If there is no link from article i to article j but $a^K_{ij} \gg a_{ij}$, then the absence of this link is considered noise, and the respective suggestion will be ranked high in the output list. In principle, the same algorithm can as well achieve the complementary task of removing unjustified existing links: if there *is* a link from i to j but $a^K_{ij} \ll a_{ij}$, then the presence of this link may be considered noise, and the link should be removed. While we intuitively anticipate our algorithm to have this capacity, we have not tested it formally.

4.6.7 Human Computation in Wikispeedia

The PCA-based method we have investigated in this chapter is purely 'intrinsic' in the sense that link suggestions are made exclusively by generalizing the information already contained in Wikipedia's link structure.

To conclude this chapter, we now sketch an 'extrinsic' method, which harnesses human computation to elicit novel information that is not necessarily implicit in the adjacency matrix yet. It has been motivated by repeated feedback from Wikispeedia players who complained that certain links they expected to find in an article were in fact missing. The idea is to make a virtue out of necessity and give frustrated players the chance to eliminate the root of their gripes, by means of a small add-on to the game of Wikispeedia.

Note that the game as described in Chapter 2 remains completely unchanged, we merely add a little box to a corner of the screen, above the actual Wikipedia page, asking, 'You think this article should link to *g*, but it doesn't?' (where *g* is the goal article of the current game), followed by a 'Report!' button (cf. Figure 2–2). When this button is clicked, the manual link suggestion is stored in the Wikispeedia database.

Obviously, confidence in a link suggestion being valuable should be higher the more frequently it has been submitted. If we assume players to be independent of each other, a suggestion can be considered reliable as soon as it has been seen twice. Another way of verifying a suggestion would be by tentatively incorporating it into the Wikipedia version used for the game and accepting it once it has actually been clicked by a number of different players, since most of them must have anticipated the link without being certain of its existence. (The same verification method could be used for predictions made by the PCA-based method.)

Since we have introduced this feature only recently (on December 9, 2009), data is still too scarce for a formal evaluation. Hence, we restrict ourselves to delivering two examples:

The player who produced the path

⟨LIBERAL DEMOCRATS, UNITED KINGDOM, TIME ZONE, **TIME**, DAY, **SUN**⟩

suggested a link from TIME to SUN, and rightfully so, since the TIME article mentions the word 'sun' twice, without containing the link to the SUN article.

In the second example,

⟨CORNEA, ROMANIA, EUROPEAN UNION, UNITED KINGDOM, **WILLIAM SHAKESPEARE**, MACBETH, **GLOBE THEATRE**⟩,

a link from WILLIAM SHAKESPEARE to GLOBE THEATRE is proposed, since clearly one would expect the WILLIAM SHAKESPEARE article to talk about (and link to) GLOBE THEATRE, the major venue in Shakespeare's lifetime. And yet it does not. So, just like the PCA-based method, this human-computation approach finds both links that can be placed in the respective article as is, as well as links that correspond to topics the article misses to mention in the first place.

4.7 Conclusion

In this chapter we have presented a novel approach to find missing links in document collections such as Wikipedia. We use exclusively the structure of Wikipedia's hyperlink graph, in a featureless approach based on principal component analysis, a mathematically sound generalization technique. It enforces the linking policy that is implicit in the entirety of Wikipedia's hyperlink structure by putting additional links into those articles that contravene the linking guidelines. The method is conceptually clean, yet its simplicity does not keep it from outperforming the state of the art.

Our method draws on work done by the commonsense reasoning community, and we strive to give an intuitive explanation of how and why it implements the paradigm of cumulative analogy by performing dimensionality reduction. We point out implications of the approach beyond link completion: it can detect topics a given Wikipedia article fails to cover, and cluster articles along semantic lines.

CHAPTER 5

Conclusion

In general, the process of collecting knowledge from people is very time-consuming and thus expensive. For example, professional lexicographers had to be employed to build WordNet [Fellbaum, 1998], and every time the database has to be extended, experts need to be consulted again. Consequently, WordNet has rather sparse coverage, with about 120,000 noun entries in version 3.0 [Miller, 2006], and is slow in responding to the emergence of new concepts. This also affects measures of semantic relatedness that are based on WordNet or similar hand-crafted ontologies, e.g., Resnik's [1999] or Rada *et al.*'s [1989].

These issues are resolved by Wikipedia, which had about 2.7 million articles as of March 2009 [Wikipedia, 2009] and is continuously kept up to date by thousands of volunteer contributors. Moreover, thanks to its hyperlink structure, Wikipedia can be considered a primitive semantic network, as we have argued in Chapter 1. However, Wikipedia has to be handled with caution in this function, since its hyperlinks represent semantic links only in a rather noisy way (cf. Section 3.3.1). One of the contributions this thesis makes is a method to filter such noise by means of a human-computation game, thus effecting a robust measure of semantic relatedness. In the next section we will recapitulate our contributions in some more detail, before concluding this thesis by pointing out directions for potential future research.

5.1 Summary of Contributions

In Chapter 2 we introduce Wikispeedia and discuss it in its function as a 'game with a purpose', i.e., as a game 'in which players perform a useful computation as a side effect of enjoyable game play' [von Ahn and Dabbish, 2008].

Chapter 3 describes in detail how the data produced by Wikispeedia players can be turned into a measure of semantic relatedness. Our approach effectively filters the noise that complicates the use of Wikipedia as a semantic network and annotates edges with importance weights. It has the advantage that we reward contributors with fun instead of money, which facilitates obtaining the data. We show that the quality of data gathered this way is sufficient for computing a reliable measure of semantic relatedness. Moreover, this approach is easily scalable: if we want to add a new concept, we simply make it the goal of some future games. The fact that the informal Wiki Game has been popular among Wikipedians for a long time means that such adaptive data collection can be achieved quickly and at low cost. Additionally, if the game traces recorded by other implementations of the game (cf. Section 2.4.2) could be obtained, they could be pooled with the Wikispeedia data to create a larger repository of game traces to learn from.

We show that the resulting relatedness measure outperforms Latent Semantic Analysis [Landauer and Dumais, 1997] when the task is to find the nearest semantic neighbors of a given concept. Although the incremental character of our measure is cognitively plausible, it may be a limitation from the practical viewpoint, in comparison to offline corpus-based methods, since we can learn the distance between two concepts only when they co-occur in a game.

To alleviate this problem, we investigate PCA-based generalization techniques, which result in our relatedness measure being defined for all pairs of Wikipedia concepts. In particular, we propose a first-order method that smoothes the matrix

of pairwise relatedness values using PCA, and a second-order method that represents concepts as rows in this smoothed matrix and defines relatedness as cosine similarity between such row vectors. Evaluating the performance of the resulting measure against a human-labeled test set, we find that this postprocessing step generalizes well, and that the second-order method yields better results than the first-order method. However, the second-order approach changes the structure of the original matrix of relatedness values drastically, such that it becomes symmetric and has close to no correlation with the original input matrix.

Our generalization methods are not specifically tailored to the relatedness matrix produced by Wikispeedia, which suggests that they can be employed to postprocess other sparse relatedness matrices, too. The fact that Wikipedia's hyperlinks bear semantic value lets us take its adjacency matrix as yet another matrix of semantic relatedness values, amenable to our PCA-based techniques. We demonstrate that, while the latter do work in this setup, too, the effected measure of semantic relatedness is less accurate than the one based on Wikispeedia data. We interpret this as confirmation of our claim that Wikispeedia is able to filter hyperlinks that do not correspond to semantic links.

Chapter 4 shows that, beyond the task of inferring semantic relatedness, running PCA on Wikipedia's adjacency matrix can also be harnessed for improving Wikipedia itself, by finding hyperlinks which are not present yet but which are suggested by the statistical structure of the ensemble of existing links. The algorithm we propose uses our first-order method directly as a subroutine. PCA is infeasible on the unmodified adjacency matrix of full Wikipedia, due to its sheer size. We must therefore reduce size beforehands. Note that this only shrinks the training set, not the set of articles for which the algorithm can predict links.

Our human user evaluation lets us conclude that, despite its conceptual simplicity, our approach outperforms the previous state of the art in the task of finding

an article's most important missing link. We complement this result by demonstrating that the order in which our algorithm ranks link suggestions is meaningful, too.

5.2 Future Directions

One main contribution of this thesis is the definition of novel measures of semantic relatedness, based on data produced by the Wikispeedia game. Our measures are untyped in the sense that they merely attribute a numerical value to a given pair of concepts. A high relatedness value just indicates that the concepts are closely related in *some* sense, without further specification. However, there are many types of semantic relatedness between concepts. Possible such types are 'is-a' (hyponymy), 'is-part-of' (meronymy), 'is-opposite-of' (antonymy), or 'is-used-for'. While the plain numerical value is sufficient for numerous applications (examples are given in Section 1.1), the exact way in which two concepts are related clearly matters a lot, especially in high-level reasoning tasks. For instance, imagine an automated shopping agent crawling the Web for bargains on behalf of a human customer, e.g., trying to find jazz records. Now, the concepts PIANO and JAZZ RECORD are closely related because oftentimes PIANO is-used-for JAZZ RECORD. Still, the agent should not purchase a piano. On the flip side, KIND OF BLUE and JAZZ RECORD are highly related, too, but in this case the agent should indeed consider the purchase, since KIND OF BLUE is-a (great) JAZZ RECORD.

Given the usefulness of typed semantic information, one direction of future research might explore the design and evaluation of human-computation games for the purpose of learning typed measures of semantic relatedness. In Section 2.4.1 we have mentioned Verbosity [von Ahn *et al.*, 2006], which collects such typed semantic information, in the form of statements such as 'PIANO is-used-for JAZZ RECORD'. We point out that the existence of one human-computation method

successfully achieving this does not imply that no further similar games should be designed, since variation is at the very heart of why people play games: they want to be entertained, and the more ways there are to find entertainment, the better.

One possible game design could be based on co-operative graph modification: Two players are given the same rudimentary semantic network, with untyped edges (relationships) and simultaneously label edges with relationship types, receiving rewards when they agree. Other actions could include the deletion or insertion of vertices (concepts). The rules would have to ensure that the data gathered is reliable and that game play is enjoyable. The edges of the input network could for instance be determined by connecting vertices (concepts) that have low Wikispeedia distance. This way, an untyped semantic net defined solely by Wikispeedia distance could incrementally be transformed into a proper, typed semantic network.

Once there are several games that collect similar types of data, it is desirable to combine the output from the different games into one single data set. For instance, data from Verbosity have been incorporated into Open Mind Common Sense (which, though not a game, is also based on human computation) [Speer, 2009]. This integration constitutes another interesting challenge.

Throughout this thesis we argue that the Wikipedia hyperlink graph can be considered a primitive semantic network and exploit this observation for computing pairwise semantic relatedness. Another line of future research could investigate to what degree the Wikipedia graph can be used more directly like typical semantic networks, which were originally devised as associative semantic memories that work through spreading activation: whenever one or more 'input' concepts are activated in the network graph, they also trigger other, neighboring concepts, and so on recursively. Such networks have been used to model human semantic memory [Quillian, 1968] and are consequently of interest for artificial intelligence research.

The fact that Wikipedia links carry semantic value would justify initializing the network with Wikipedia's raw hyperlink structure, all edges having equal weight. These weights could then be fine-tuned according to typically occurring activity patterns, possibly online, as the semantic network is deployed. Hopfield networks [Hopfield, 1982], recurrent neural nets that can serve as associative memories, could be an apt framework to achieve this. With roughly 2.7 million concepts, the coverage of such a semantic network would be unprecedented. Because of the sheer size, weight learning would be challenging in its own right and would possibly call for novel algorithms.

Finally, future research should develop real-world applications to harness the measures of semantic relatedness that have been defined using Wikipedia. Although many domains such as natural-language processing, information retrieval, and human–computer interaction could profit this way, 'So far the algorithms [...] are underutilized, given the large advances in accuracy and vocabulary that they offer,' according to Medelyan *et al.* [2009].

It might even be possible that in terms of accuracy not much improvement is possible over the existing approaches. No matter how sophisticated the algorithm, when assessing the performance of computational methods for inferring semantic relatedness one should always bear in mind that we will never be able to attain perfection, for the very notion of semantic relatedness is only weakly defined: low inter-human correlations show that, even amongst those who are providing the 'ground truth', overwhelming agreement is lacking.—But is it not exactly such ambiguity that makes human language challenging and fun? Why else would people enjoy the version of our opening example that says:

'Time flies like an arrow. Fruit flies like a banana.'

REFERENCES

[Amazon, 2009] Amazon. Amazon Mechanical Turk. Website, 2009. `http://www.mturk.com` (accessed Feb. 9, 2010).

[Apache, 2009] Apache. Lucene. Website, 2009. `http://lucene.apache.org` (accessed Feb. 9, 2010).

[Brin and Page, 1998] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998.

[Budanitsky and Hirst, 2001] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proc. Workshop on WordNet and Other Lexical Resources*, 2001.

[Cates, 2009] A. Cates. SOS Children's Villages UK. Personal communication, 2009.

[Chklovski, 2003] T. Chklovski. Learner: A system for acquiring commonsense knowledge by analogy. In *Proc. 2nd International Conference on Knowledge Capture (K-CAP-03)*, 2003.

[Cilibrasi and Vitányi, 2007] R. L. Cilibrasi and P. Vitányi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

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

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

[Finkelstein *et al.*, 2002] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

[Fissaha Adafre and de Rijke, 2005] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Proc. 3rd International Workshop on Link Discovery (LinkKDD-05)*, 2005.

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

Analysis. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.

[Gabrilovich, 2002] E. Gabrilovich. WordSimilarity-353 test collection. Website, 2002. `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html` (accessed Feb. 8, 2010).

[Google, 2010] Google. Google Image Labeler. Website, 2010. `http://images.google.com/imagelabeler` (accessed Feb. 9, 2010).

[He *et al.*, 2007] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the deep Web. *Communications of the ACM*, 50(5):94–101, 2007.

[Hinton and Salakhutdinov, 2006] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[Hopfield, 1982] J. J. Hopfield. Neural networks and physical systems with emergent collective computational properties. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[Huettig *et al.*, 2006] F. Huettig, P. T. Quinlan, S. A. McDonald, and G. T. M. Altmann. Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121(1):65–80, 2006.

[Kaur and Hornof, 2005] I. Kaur and A. J. Hornof. A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI-05)*, 2005.

[Landauer and Dumais, 1997] T. Landauer and S. T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[Landauer and Kintsch, 1998] T. Landauer and W. Kintsch. LSA. Website, 1998. `http://lsa.colorado.edu` (accessed Feb. 8, 2010).

[Landauer *et al.*, 1998] T. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

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

[Manning *et al.*, 2008] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Medelyan *et al.*, 2009] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from Wikipedia. *International Journal of Human–Computer Studies*, 67(9):716–754, 2009.

[Mihalcea and Csomai, 2007] R. Mihalcea and A. Csomai. Wikify! Linking documents to encyclopedic knowledge. In *Proc. 16th ACM Conference on Information and Knowledge Management (CIKM-07)*, 2007.

[Miller, 2006] G. A. Miller. WordNet 3.0 documentation. Website, 2006. `http://wordnet.princeton.edu/wordnet/documentation` (accessed Feb. 1, 2010).

[Milne and Witten, 2008a] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. 1st AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI-08)*, 2008.

[Milne and Witten, 2008b] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proc. 17th ACM Conference on Information and Knowledge Management (CIKM-08)*, 2008.

[Milne, 2007] D. Milne. Computing semantic relatedness using Wikipedia link structure. In *Proc. New Zealand Computer Science Research Student Conference*, 2007.

[Milne, 2009] D. Milne. WikipediaMiner toolkit. Website, 2009. `http://wikipedia-miner.sourceforge.net` (accessed June 6, 2009).

[Ollivier and Senellart, 2007] Y. Ollivier and P. Senellart. Finding related pages using Green measures: An illustration with Wikipedia. In *Proc. 22nd AAAI Conference on Artificial Intelligence (AAAI-07)*, 2007.

[Pearson, 1901] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[Quillian, 1968] M. R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*. MIT Press, 1968.

[Rada *et al.*, 1989] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[Resnik, 1999] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[Singh *et al.*, 2002] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open Mind Common Sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA,*

and ODBASE, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer Verlag, 2002.

[Snow *et al.*, 2008] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, 2008.

[Speer *et al.*, 2008] R. Speer, C. Havasi, and H. Lieberman. AnalogySpace: Reducing the dimensionality of common sense knowledge. In *Proc. 23rd AAAI Conference on Artificial Intelligence (AAAI-08)*, 2008.

[Speer, 2009] R. Speer. Open Mind Common Sense blog. Website, 2009. `http://conceptnet.blogspot.com/2009/08/verbosity-and-one-meeeelion-sentences.html` (accessed Feb. 6, 2010).

[Strube and Ponzetto, 2006] M. Strube and S. P. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. 21st National Conference on Artificial Intelligence (AAAI-06)*, 2006.

[Turk and Pentland, 1991] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Turney, 2001] P. D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. 12th European Conference on Machine Learning (ECML-01)*, 2001.

[Tversky, 1977] A. Tversky. Features of similarity. *Psychological Review*, 84(2):327–352, 1977.

[Veksler *et al.*, 2008] V. D. Veksler, R. Z. Govostes, and W. D. Gray. Defining the dimensions of the human semantic space. In *Proc. Conference of the Cognitive Science Society (CogSci-08)*, 2008.

[von Ahn and Dabbish, 2004] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI-04)*, 2004.

[von Ahn and Dabbish, 2008] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[von Ahn *et al.*, 2006] L. von Ahn, M. Kedia, and M. Blum. Verbosity: A game for collecting common-sense facts. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI-06)*, 2006.

[West *et al.*, 2009a] R. West, J. Pineau, and D. Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *Proc. 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, 2009.

[West *et al.*, 2009b] R. West, D. Precup, and J. Pineau. Completing Wikipedia's hyperlink structure through dimensionality reduction. In *Proc. 18th ACM Conference on Information and Knowledge Management (CIKM-09)*, 2009.

[West, 2009a] R. West. Project website, 2009. `http://www.cs.mcgill.ca/~rwest/link-suggestion` (accessed Feb. 9, 2010).

[West, 2009b] R. West. Wikispeedia. Website, 2009. `http://www.wikispeedia.net` (accessed Feb. 9, 2010).

[Wikipedia Game, 2010] Wikipedia Game. Website, 2010. `http://www.wikipediagame.org` (accessed Jan. 4, 2010).

[Wikipedia Maze, 2010] Wikipedia Maze. Website, 2010. `http://www.wikipediamaze.com` (accessed Jan. 4, 2010).

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

[Wikipedia, 2008] Wikipedia. 2008/9 Wikipedia Selection for schools. Website, 2008. `http://schools-wikipedia.org` (accessed June 3, 2009).

[Wikipedia, 2009] Wikipedia. Data dump of March 6, 2009. Website, 2009. `http://download.wikimedia.org/enwiki/20090306` (accessed June 3, 2009).

[Wikipedia, 2010a] Wikipedia. Help:Infobox. Website, 2010. `http://en.wikipedia.org/w/index.php?title=Help:Infobox&oldid=335883227` (accessed Feb. 5, 2010).

[Wikipedia, 2010b] Wikipedia. Wikipedia:Linking. Website, 2010. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Linking&oldid=342061829` (accessed Feb. 5, 2010).

[Wikipedia, 2010c] Wikipedia. Wikipedia:Manual of Style. Website, 2010. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style&oldid=342148400#Wikilinks` (accessed Feb. 5, 2010).

[Wikipedia, 2010d] Wikipedia. Wikipedia:Wiki Game. Website, 2010. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Wiki_Game&oldid=340765784` (accessed Feb. 5, 2010).

[Wikirace, 2010] Wikirace. Website, 2010. `http://www.wikirace.org` (accessed Jan. 4, 2010).

[Wu and Weld, 2007] F. Wu and D. S. Weld. Autonomously semantifying Wikipedia. In *Proc. 16th ACM Conference on Information and Knowledge Management (CIKM-07)*, 2007.

APPENDIX A

Mathematical Notation

Logarithms. All logarithms are binary, i.e., $\log x := \log_2 x$.

Sets. Sets are denoted by capital Fraktur letters: $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$, $\mathfrak{D}$, etc. Set cardinality is designated by vertical bars; e.g., if $\mathfrak{A} = \{x_1, x_2, ..., x_n\}$, then $|\mathfrak{A}| := n$.

Matrices and Vectors. Matrices are denoted by bold capital letters: **A**, **B**, **C**, **D**, etc.

Row vectors of a matrix are referred to by small bold letters, followed by their row index; e.g., $\mathbf{a}_i$ is the i-th row of matrix **A**.

Entries of a matrix are denoted by small regular letters, followed by their row and column indices; e.g., a_{ij} is the entry in row i and column j of matrix **A**.

To define the entries of a matrix, we sometimes use parenthesis notation; e.g., $\mathbf{A} = (i + j)$ means $a_{ij} = i + j$.

The transpose of a matrix **A** is denoted by $\mathbf{A}^{\mathrm{T}}$.

The ℓ_2-norm of a vector $\mathbf{a}_i$ is written as $\|\mathbf{a}_i\|$.

APPENDIX B

Wikipedia Article about STATISTICS

On the following pages we reproduce the article about STATISTICS from the 2008/9 Wikipedia Selection for schools [Wikipedia, 2008]. Hyperlinks to other articles are underlined. The links that have been introduced by our hyperlink suggestion algorithm (cf. Table 4–2) are marked with a star (★).

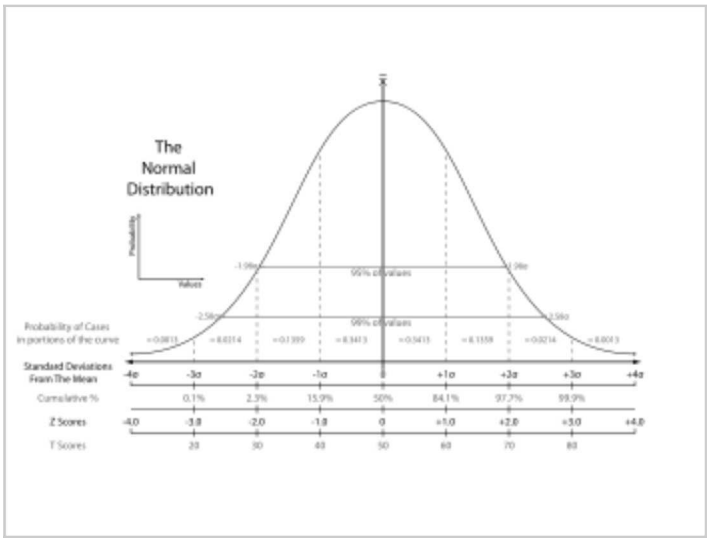
Statistics

2008/9 Schools Wikipedia Selection. Related subjects: Mathematics

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. It is applicable to a wide variety of academic disciplines, from the natural and social sciences to the humanities, and to government and business.

Statistical methods can be used to summarize or describe a collection of data; this is called **descriptive statistics**. In addition, patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, and then used to draw inferences about the process or population being studied; this is called **inferential statistics**. Both descriptive and inferential statistics comprise **applied statistics**. There is also a discipline called **mathematical statistics**, which is concerned with the theoretical basis of the subject.

The word ***statistics*** is also the plural of ***statistic*** (singular), which refers to the result of applying a statistical algorithm to a set of data, as in economic statistics, crime statistics, etc.



A graph of a normal bell curve showing statistics used in standardized testing assessment. The scales include *standard deviations, cumulative percentages, percentile equivalents, Z-scores, T-scores, standard nines,* and *percentages in standard nines.*

History

Statistics arose, no later than the 18th century, from the need of states to collect data on their people and economies, in order to administer them. Its meaning broadened in the early 19th century to include the collection and analysis of data in general. Today statistics is widely employed in government, business, and the natural and social sciences.

Because of its origins in government and its data-centric world view, statistics is considered to be not a subfield of mathematics but rather a distinct field that uses mathematics. Its mathematical foundations were laid in the 17th and 18th centuries with the development of probability theory. The method of least squares, a central technique of the discipline, was invented in the early 19th century by several authors. Since then new techniques of probability and statistics have been in continual development. Modern computers have expedited large-scale statistical computation, and have also made possible new methods that would be impractical to perform manually.

Overview

In applying statistics to a scientific, industrial, or societal problem, one begins with a process or population to be studied. This might be a population of people in a country, of crystal grains in a rock, or of goods manufactured by a particular factory during a given period. It may instead be a process observed at various times; data collected about this kind of "population" constitute what is called a time series.

For practical reasons, rather than compiling data about an entire population, one usually studies a chosen subset of the population, called a sample. Data are collected about the sample in an observational or experimental setting. The data are then subjected to statistical analysis, which serves two related purposes: description and inference.

- Descriptive statistics can be used to summarize the data, either numerically or graphically, to describe the sample. Basic examples of numerical descriptors include the mean and standard deviation. Graphical summarizations include various kinds of charts and graphs.
- Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences about the larger population. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation), descriptions of association (correlation), or modeling of relationships (regression). Other modeling techniques include ANOVA, time series, and data mining.

> "... it is only the manipulation of uncertainty that interests us. We are not concerned with the matter that is uncertain. Thus we do not study the mechanism of rain; only whether it will rain."
> Dennis Lindley, "The Philosophy of Statistics", *The Statistician* (2000).

The concept of correlation is particularly noteworthy. Statistical analysis of a data set may reveal that two variables (that is, two properties of the population under consideration) tend to vary together, as if they are connected. For example, a study of annual income and age of death among people might find that poor people tend to have shorter lives than affluent people. The two variables are said to be correlated. However, one cannot immediately infer the existence of a causal relationship between the two variables (see Correlation does not imply causation). The correlated phenomena could be caused by a third, previously unconsidered phenomenon, called a lurking variable.

If the sample is representative of the population, then inferences and conclusions made from the sample can be extended to the population as a whole. A major problem lies in determining the extent to which the chosen sample is representative. Statistics offers methods to estimate and correct for randomness in the sample and in the data collection procedure, as well as methods for designing robust experiments in the first place (see experimental design).

The fundamental mathematical concept employed in understanding such randomness is probability. Mathematical statistics (also called statistical theory) is the branch of applied mathematics that uses probability theory and analysis to examine the theoretical basis of statistics.

The use of any statistical method is valid only when the system or population under consideration satisfies the basic mathematical assumptions of the method. Misuse of statistics can produce subtle but serious errors in description and interpretation — subtle in that even experienced professionals sometimes make such errors, and serious in that they may affect social policy, medical practice and the reliability of structures such

as bridges and nuclear power plants. Even when statistics is correctly applied, the results can be difficult to interpret for a non-expert. For example, the statistical significance of a trend in the data — which measures the extent to which the trend could be caused by random variation in the sample — may not agree with one's intuitive sense of its significance. The set of basic statistical skills (and skepticism) needed by people to deal with information in their everyday lives is referred to as statistical literacy.

Statistical methods

Experimental and observational studies

A common goal for a statistical research project is to investigate causality, and in particular to draw a conclusion on the effect of changes in the values of predictors or independent variables on response or dependent variables. There are two major types of causal statistical studies, experimental studies and observational studies. In both types of studies, the effect of differences of an independent variable (or variables) on the behaviour of the dependent variable are observed. The difference between the two types is in how the study is actually conducted. Each can be very effective.

An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation may have modified the values of the measurements. In contrast, an observational study does not involve experimental manipulation. Instead data are gathered and correlations between predictors and the response are investigated.

An example of an experimental study is the famous Hawthorne studies which attempted to test changes to the working environment at the Hawthorne plant of the Western Electric Company. The researchers were interested in whether increased illumination would increase the productivity of the assembly line workers. The researchers first measured productivity in the plant then modified the illumination in an area of the plant to see if changes in illumination would affect productivity. As it turns out, productivity improved under all the experimental conditions (see Hawthorne effect). However, the study is today heavily criticized for errors in experimental procedures, specifically the lack of a control group and blindedness.

An example of an observational study is a study which explores the correlation between smoking and lung cancer. This type of study typically uses a survey to collect observations about the area of interest and then perform statistical analysis. In this case, the researchers would collect observations of both smokers and non-smokers, perhaps through a case-control study, and then look at the number of cases of lung cancer in each group.

The basic steps for an experiment are to:

1. plan the research including determining information sources, research subject selection, and ethical considerations for the proposed research and method,
2. design the experiment concentrating on the system model and the interaction of independent and dependent variables,
3. summarize a collection of observations to feature their commonality by suppressing details (descriptive statistics),
4. reach consensus about what the observations tell us about the world we observe (statistical inference),
5. document and present the results of the study.

Levels of measurement

> *See: Stanley Stevens' "Scales of measurement" (1946): nominal, ordinal, interval, ratio*

There are four types of measurements or measurement scales used in statistics. The four types or levels of measurement (nominal, ordinal, interval, and ratio) have different degrees of usefulness in statistical research. Ratio measurements, where both a zero value and distances between different measurements are defined, provide the greatest flexibility in statistical methods that can be used for analyzing the data. Interval measurements have meaningful distances between measurements but no meaningful zero value (such as IQ measurements or temperature measurements in Fahrenheit). Ordinal measurements have imprecise differences between consecutive values but a meaningful order to those values. Nominal measurements have no meaningful rank order among values.

Variables conforming only to nominal or ordinal measurements are together sometimes called categorical variables, since they cannot reasonably be numerically measured, whereas ratio and interval measurements are grouped together as quantitative or continuous variables due to their numerical nature.

Statistical techniques

Some well known statistical tests and procedures for research observations are:

- Student's t-test
- chi-square test
- Analysis of variance (★) (ANOVA)
- Mann-Whitney U
- Regression analysis
- Factor Analysis
- Correlation
- Pearson product-moment correlation coefficient
- Spearman's rank correlation coefficient
- Time Series Analysis

Specialized disciplines

Some fields of inquiry use applied statistics so extensively that they have specialized terminology. These disciplines include:

- Actuarial science
- Applied information economics (★)
- Biostatistics
- Bootstrap & Jackknife Resampling
- Business statistics
- Data mining (applying statistics and pattern recognition to discover knowledge from data)
- Demography
- Economic statistics (Econometrics)
- Energy statistics
- Engineering (★) statistics
- Environmental Statistics
- Epidemiology
- Geography and Geographic Information Systems, more specifically in Spatial analysis
- Image processing

- Multivariate Analysis
- Psychological statistics
- Quality
- Social statistics
- Statistical literacy
- Statistical modeling
- Statistical surveys
- Process analysis and chemometrics (for analysis of data from analytical chemistry and chemical engineering)
- Survival analysis
- Reliability engineering
- Statistics in various sports, particularly baseball and cricket

Statistics form a key basis tool in business and manufacturing as well. It is used to understand measurement systems variability, control processes (as in statistical process control or SPC), for summarizing data, and to make data-driven decisions. In these roles it is a key tool, and perhaps the only reliable tool.

Statistical computing

The rapid and sustained increases in computing power starting from the second half of the 20th century have had a substantial impact on the practice of statistical science. Early statistical models were almost always from the class of linear models, but powerful computers, coupled with suitable numerical algorithms, caused a resurgence of interest in nonlinear models (especially neural networks and decision trees) and the creation of new types, such as generalised linear models and multilevel models.

Increased computing power has also led to the growing popularity of computationally-intensive methods based on resampling, such as permutation tests and the bootstrap, while techniques such as Gibbs sampling have made Bayesian methods more feasible. The computer revolution has implications for the future of statistics, with a new emphasis on "experimental" and "empirical" statistics. A large number of both general and special purpose statistical packages are now available to practitioners.

Misuse

There is a general perception that statistical knowledge is all-too-frequently intentionally misused, by finding ways to interpret the data that are favorable to the presenter. A famous saying attributed to Benjamin Disraeli is, " There are three kinds of lies: lies, damned lies, and statistics." And Harvard President Lawrence Lowell wrote in 1909 that statistics, "like veal pies, are good if you know the person that made them, and are sure of the ingredients."

If various studies appear to contradict one another, then the public may come to distrust such studies. For example, one study may suggest that a given diet or activity raises blood pressure, while another may suggest that it lowers blood pressure. The discrepancy can arise from subtle variations in experimental design, such as differences in the patient groups or research protocols, that are not easily understood by the non-expert. (Media reports sometimes omit this vital contextual information entirely.)

By choosing (or rejecting, or modifying) a certain sample, results can be manipulated; throwing out outliers is one means of doing so. Such manipulations need not be malicious or devious; they can arise from unintentional biases of the researcher. The graphs used to summarize data can also be misleading.

Deeper criticisms come from the fact that the hypothesis testing approach, widely used and in many cases required by law or regulation, forces one hypothesis (the null hypothesis) to be "favored", and can also seem to exaggerate the importance of minor differences in large studies. A difference that is highly statistically significant can still be of no practical significance. (See criticism of hypothesis testing and controversy over the null hypothesis.)

One response has been a greater emphasis on the *p*-value over simply reporting whether a hypothesis was rejected at the given level of significance. The *p*-value, however, does not indicate the size of the effect. Another increasingly common approach is to report confidence intervals. Although these are produced from the same calculations as hypothesis tests or *p*-values, they describe both the size of the effect and the uncertainty surrounding it.