

SEMANTICS EXTRACTION IN INFORMATION SPACES USING CO-OCCURRENCE ANALYSIS

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

by

MANDAR R. MUTALIKDESAI



International Institute of Information Technology – Bangalore

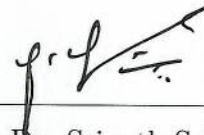
2013

Dedicated to
Amma, Pappa
and
Kshithi

Certificate

This is to certify that the thesis titled "Semantics Extraction in Information Spaces using Co-occurrence Analysis" being submitted by Mandar R. Mutalikdesai to the International Institute of Information Technology – Bangalore, for the award of the degree of Doctor of Philosophy, is a record of bona fide research work carried out by him under my supervision, and Mandar R. Mutalikdesai fulfills the requirements of the regulations of the degree. The results in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Date: 02 March 2013



Dr. Srinath Srinivasa
Associate Professor

International Institute of Information Technology – Bangalore

INTERNATIONAL INSTITUTE OF INFORMATION
TECHNOLOGY, BANGALORE.
India
26 / C, Electronics City,
Hosur Road, Bangalore - 560 100



Declaration

This is to certify that the thesis titled "Semantics Extraction in Information Spaces using Co-occurrence Analysis" being submitted to the International Institute of Information Technology – Bangalore, for the award of the degree of Doctor of Philosophy, is a record of bonafide research work carried out by me. The results in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Date: 02 March 2013



Mandar R. Mutalikdesai
International Institute of Information Technology – Bangalore
India

Acknowledgements

I would like to begin by expressing my gratitude towards Prof. Srinath Srinivasa, my thesis supervisor. I have had the pleasure of working with Prof. Srinath for the last 10 years (spanning my Masters and PhD programs), during which I have learned a lot from him. He has always afforded me the freedom to try different things (and to try things differently!), even if it meant that I would not always succeed. I especially thank him for providing me this freedom. That is how I have been able to learn. And of course, I am grateful to him for his timely advice and contributions during all stages of this work. I will cherish the many discussions I have had with him on technical and philosophical matters. I hope to collaborate with him in the future as well.

I received some very insightful and constructive comments from the four anonymous reviewers of this thesis. Their comments and ideas greatly helped in the shaping of this thesis. For this, I am very grateful to them.

All the years that I have been at IIIT-B, I have been a part of the Open Systems Lab. Over the years, I have had the pleasure of working with many lab-mates. I would like to thank Sanket Patil, Aditya Rachakonda, Karthik B. R., Siddhartha Reddy, Sumant Kulkarni, Sudha Mani, Rashmi Rao, Vijay Olety, Nikhil Patil and Saikat Mukherjee for the various vibrant discussions we had in the lab.

I would like to express the deepest gratitude to Prof. S. Sadagopan for providing a vibrant atmosphere for learning and research at IIIT-B. I would also like to thank the faculty members of IIIT-B, many of whom have provided insightful suggestions on my thesis topic. I am especially grateful to Prof. Rajagopalan, Prof. Chandrashekar and Prof. Prasanna, whose insightful comments vastly improved the ideas in this thesis. Thanks are also due to Mr. Ramachandra A. N., Mr. C. M. Abraham, Mr. Murugan, Ms. Chandrika, Ms. Rama, Ms. Padma, Ms. Nirmala and Mr. Somashekhar for smoothly handling all administrative issues during my stay at IIIT-B.

Special thanks to Prof. Shalini Urs for her constant support and encouragement. I have enjoyed working with her at ISiM (UoM, Mysore) over the

last 4 years, and hope that we will continue to work together in the future as well. I would also like to convey my gratitude to Prof. S. K. Gupta of IIT-Delhi for his analytical comments on this thesis.

For the first three years of my PhD program, I was an Infosys Scholar. I would like to convey my gratitude to Infosys Ltd. for their encouragement and support.

I have had the pleasure of collaborating with Prof. Ambuj K. Singh and Prof. B. S. Manjunath during my stay at the University of California, Santa Barbara. I would like to thank them for the many discussions I have had with them during the formative stages of my PhD work.

My friends and extended family have always been very supportive of what I have wanted to do in my life and career. My thanks to Vijay, Udaka, Raghu, Deepak, Archana, Pawan, Alpana, Payal, Prashant, Daya, Kshama, Pradeep, Sanket, Surabhi, Dattu, Gopi, Chitra, Sudarsan, Kavan, Aniruddha, Prasad Joshi, Prasad Kalghatgi, Shyam, Ram, Manoj Deshpande, Manoj Markod, Satish, Narasimha, Shailesh, Arvind, Sunita, Raghavendra Edke, Harshad and Shrinidhi.

I would like to convey special gratitude to my little nephew, Naman, for being such a lovely part of my life.

Words cannot express what I feel towards my parents, Asha – my sister, my mother-in-law and Kshithi – my closest friend. I possibly couldn't express enough gratitude to them for their endless love, care and inspiration.

Abstract

An information space is a large collection of content generated by human actors. Examples of information spaces include the Web, digital libraries and social media. These spaces hold significant amounts of latent semantics, which may be relevant to various stakeholders such as market data analysts, marketing research personnel, etc. In this thesis, we address the problem of mining latent semantics from information spaces. We note that information spaces are not uniformly similar in nature. They can be classified into two types: *repository spaces* and *social spaces*. Examples of repository spaces are the Web and digital libraries, while examples of social spaces are the blogosphere and wikis. Content is generated in both these spaces by *cognitive processes* of actors, with the content manifesting as documents, web pages, blog posts, reviews, articles, tweets, etc. During the creation of such content, the actor typically embeds her *individual world-views* (opinions, feelings, etc.) into the content. Also, in both these spaces, there exist social interactions between the cognitive processes, which manifest as references between documents in the form of links or citations, comments to blog posts, rebuttals to criticisms, responses to reviews, etc. However, repository spaces and social spaces differ in terms of the boundary of social interaction, the scope of engagement of actors, and the localized synchrony of social interactions.

While the social interactions in repository spaces can be spread across the entire repository space, the social interactions in social spaces have well-defined boundaries within which the *commonly held world-views* of multiple actors can emerge in a focused manner. We call this boundary within which cognitive processes interact as a *socio-cognitive process*. While the entire repository represents a single socio-cognitive process, there exist multiple socio-cognitive processes within a social space. Also, in a repository space, the engagement of actors is typically limited to editing only a few documents, since they can edit only those documents that are owned by them. In social spaces, on the other hand, actors can engage in multiple socio-cognitive processes, typically even if those socio-cognitive processes are not owned by them. Another aspect in which repository spaces and social spaces differ is

the localized synchrony of social interactions. As we have observed, the social interactions in a repository space are not localized. Also, these interactions between cognitive processes do not exhibit synchrony. By this, we mean that these social interactions are not in tune with each other over time within the socio-cognitive process. On the other hand, the social interactions in a social space are not only localized to a well-defined socio-cognitive process, but also synchronous. Such social interactions largely take place within short durations of each other in a localized manner.

In the first part of this thesis, we posit that *semantics* are the commonly held world-views of actors, which emerge in a socio-cognitive process in both, repository spaces as well as social spaces. We rely on the *co-occurrence analysis* of artifacts such as concepts (e.g., named entities) and social interactions (e.g., citations) within an information space in order to mine semantics. We assert that co-occurrence analysis embodies a fundamental principle of human cognition known as the Hebbian Theory. We also assert that co-occurrence analysis is a manifestation of the principles of Ordinary Language Philosophy, which states that the meaning of a term depends upon its usage with other terms. Further, we argue that co-occurrence analysis not only helps in identifying the meaning of a concept, but also its *semantic associations* relative to other concepts. We test this hypothesis in both, repositories and social spaces.

In the second part of this thesis, we analyze the co-occurrences of citations (or co-citations) to discover *endorsed citations* in a repository space. Given a document, an endorsed citation is an outgoing citation whose target document is deemed by other co-citing documents to be more relevant to the source document than the targets of other outgoing citations from the source document. We envisage the use of endorsed citations in focused resource discovery and relevance ranking in repository spaces. In the third part of this thesis, we analyze the co-occurrences of concepts (terms) in a social space to detect *object-attribute relationships*. Given an object (i.e. a concept), we define as its attributes those concepts that, besides being semantically related to the object, help in *collectively* describing the object *uniquely*. We assert that the attributes of an object tend to co-occur with the object across

cognitive contexts (paragraphs, article-sections, documents, etc.) in a social space. We present two co-occurrence based hypotheses for identifying object-attribute relationships between concepts. We envisage that the discovery of the semantic attributes of an object has applications in social media analytics, e.g., (i) marketing research personnel looking to find out how the population characterizes their product, and (ii) classification of concepts within an encyclopedic environment like Wikipedia.

Contents

1	Introduction	1
1.1	Information Spaces as Cognitive Spaces	3
1.2	Repository Spaces vs Social Spaces	5
1.2.1	Boundary of Social Interaction	5
1.2.2	Scope of Actor Engagement	6
1.2.3	Localized Synchrony of Social Interactions	7
1.3	Semantics in Information Spaces	10
1.3.1	Repository Spaces	11
1.3.2	Social Spaces	13
1.4	Mapping the Problem Domain	15
1.4.1	Pain Points	17
1.5	Contributions of this Thesis	20
1.6	Organization of this Thesis	21
2	Related Literature	23
2.1	Models for Semantics Extraction	24

2.1.1	Existing Approaches for Semantics Extraction	25
2.1.2	Co-occurrence Analysis for Semantics Extraction	30
2.2	Co-citation Analysis in Scientific Literature and the Web	32
2.3	Detection of Object-Attribute Relationships	36
2.3.1	Attribute Labeling	37
2.3.2	Ontology Learning and Building	37
2.3.3	Attribute Detection of Actors in Social Media	38
2.3.4	Commonsense Knowledge Acquisition	39
2.3.5	Relationship Mining	39
3	Co-occurrence based Semantics Mining	42
3.1	Ordinary Language Philosophy and Co-occurrence	43
3.2	Hebbian Learning and Co-occurrence	44
3.3	Beyond OLP and Hebbian Theory	45
4	Co-citation Analysis in Repositories	47
4.1	Interpretations of Co-citations	48
4.1.1	Endorsements of Citations	49
4.1.2	Knowledge Aggregation	51
4.1.3	Conditional Relevance	53
4.2	Co-citations as Citation Endorsements	56
4.2.1	Endorsed Citation Graph	58
4.2.2	Experimental Analyses	60
4.3	Document Ranking using Endorsed Citations	76
4.3.1	Formalization of ERank	77
4.3.2	Experimental Analysis	80

5	Attribute Detection in Social Spaces	84
5.1	Co-occurrence based Attribute Detection	87
5.1.1	Modeling Co-occurrence Patterns of Concepts	88
5.1.2	The Notion of Attribute Detection	91
5.2	The Usability Hypothesis	92
5.2.1	Usability Ranking	94
5.2.2	Object-Attribute Trees	95
5.2.3	Experimental Analysis	97
5.3	The Positive Relevance Hypothesis	103
5.3.1	Quantifying Positive Relevance	104
5.3.2	Experimental Analysis	104
5.4	Usability Ranking vs Positive Relevance	105
6	Concluding Remarks	110
6.1	Limitations of our Work	112
6.1.1	Citation Endorsement	113
6.1.2	Attribute Detection	114
6.2	Future Work	116
6.2.1	Short-term Directions for Future Work	116
6.2.2	Long-term Directions for Future Work	118
A	NP-Hardness of Determinability	122
A.1	Intractability of Determinability	123
A.2	Determinability is not in NP	124
B	List of Related Publications	129

List of Figures

1.1	Illustration of the differences between repository spaces and social spaces	10
4.1	The process of a co-citation endorsing a citation	50
4.2	A co-citation as a knowledge aggregation activity	52
4.3	Co-citations as indicators of conditional relevance. The relevance of B to A can be given by $\frac{ X \cap Y }{ X }$	54
4.4	Sample citation graph	57
4.5	Pareto cumulative distribution (log-scale) of co-citation counts for endorsed citations in the Web crawl	63
4.6	Pareto cumulative distribution (log-scale) of indegrees for the Web crawl ECG	65
4.7	Pareto cumulative distribution (log-scale) of outdegrees for the Web crawl ECG	66
4.8	Pareto cumulative distribution (log-scale) of component sizes in the Web crawl ECG	67

4.9	Pareto cumulative distribution (log-scale) of co-citation counts for endorsed citations in the CiteSeer snapshot	68
4.10	Pareto cumulative distribution (log-scale) of indegrees for the CiteSeer ECG	70
4.11	Pareto cumulative distribution (log-scale) of outdegrees for the CiteSeer ECG	71
4.12	Pareto cumulative distribution (log-scale) of component sizes in the CiteSeer ECG	72
4.13	Some interesting structural motifs in the ECGs	72
4.14	A component of the Web crawl ECG showing bridging of different dense clusters	75
5.1	A hypothetical OAT in the semantic context of <i>Barack Obama</i>	96
5.2	UseRank: Frequency distribution of the a_q scores (rounded off to the nearest integer) for the 30 trial queries	101
5.3	UseRank: p_q scores for the root concepts for the 30 trial queries. The X-axis represents the trial queries, and the Y-axis represents the p_q scores.	102
5.4	UseRank: Frequency distribution of the p_q scores (rounded off to one decimal place) for the 30 trial queries	102
5.5	Positive Relevance: p_q scores for the 30 trial queries. The X-axis represents the trial queries, and the Y-axis represents the p_q scores.	106
5.6	Positive Relevance: Frequency distribution of the p_q scores (rounded off to one decimal place) for the 30 trial queries	106

5.7	Consolidated view of the p_q scores for 23 queries “common” to the two proposed approaches. The X-axis represents the queries, and the Y-axis represents the p_q scores.	107
5.8	Illustration of the similarity (in terms of the Jaccard Coefficient) between the top-10 attributes generated by the two proposed approaches for the “common” queries. The X-axis represents the queries, and the Y-axis represents the Jaccard Coefficients.	108
A.1	Illustration of the poset $(2^{N(x)} \subseteq)$	125
A.2	Illustration of <i>increase</i> in determinability in a superset of A	125
A.3	Illustration of <i>decrease</i> in determinability in a superset of B	126

List of Tables

1.1	Semantics extraction in information spaces: The problem domain	17
4.1	Co-citations of document <i>A</i> with its out-neighbors and the corresponding citation-endorsement probabilities	57
5.1	List of queries used for evaluation of the usability ranking approach. Queries 1, 7, 9, 15, 16, 26 and 30 had root concepts different from the query. The root concepts of these queries have been mentioned in parentheses.	100

1

Introduction

An information space is essentially a large collection of content generated by various human actors (users). Enormous amounts of data are generated daily in online information spaces like digital libraries, intranets, the Web, forums, blogs and wikis. Information spaces hold significant amounts of latent knowledge that may give insights to various stakeholders. For instance, an analyst in a technology product company (e.g., Apple Inc.) would be interested in knowing what the population at large opines about its latest

product, on technology forums such as Slashdot.¹ Such an analyst would be interested in questions such as: (i) what qualifiers are people assigning (*expensive*, *classy*, etc.) to our product, (ii) which celebrities are talking about our product, etc. Hence, analytics of information spaces has been attracting increasing research attention.

There exist several approaches for semantics extraction in information spaces, such as dimensionality reduction (e.g., [Deerwester et al., 1990; Song and Park, 2007; Steyvers and Griffiths, 2007]), machine learning (e.g., [Harish et al., 2010; Pang et al., 2002; Sebastiani, 2002; Turney, 2001]), generative models (e.g., [Anthes, 2010; Blei et al., 2003; Hofmann, 1999a]), and network models (e.g., [Ceglowski et al., 2003; Iria et al., 2007; Jin et al., 2009; Minkov et al., 2006; Rachakonda and Srinivasa, 2006]). Usually, these approaches view an information space as a corpus of text or hypertext, and look for patterns within the text and linkage structure. However, they do not seem to present a theoretical understanding of how semantics come to be embedded in the information space, in the first place. In comparison, we look to address this by understanding what causes content to be created by actors in information spaces, and what do semantics embedded in such content represent.

In this thesis, we classify information spaces into two types: (i) *Repository Spaces*, and (ii) *Social Spaces*. Examples of repository spaces include the Web, digital libraries, and text corpora, while examples of social spaces include blogging environments, microblogs and other social networks, and online document tagging systems.

¹<http://slashdot.org/>

1.1 Information Spaces as Cognitive Spaces

Content is created in both, repository spaces as well as social spaces, through the execution of “cognitive processes” by actors. Given an actor creating content on a certain topic in information spaces, we define her *cognitive process* as the process by which she embeds her “world-view” of the topic into the content. In other words, the actor’s cognitive process (CP) causes the content to capture what she “thinks”, “feels”, “knows” or “learns” about the topic in question. Examples of CPs include essays, monologues, argumentations, comments, rants, reviews, etc.

Suppose an actor writes a review of the Apple iPhone 4S (on a web page or in a blog post, for instance). The content of this review reflects how the actor cognitively characterizes the topic of “Apple iPhone 4S”. Suppose she characterizes the iPhone as: (i) being a smart-phone, (ii) replaceable by the Samsung Android phone, (iii) containing the iOS operating system, etc. These characterizations reflect the world-view that the actor holds about the iPhone. For example, she thinks that:

- *iPhone* “is a” *smart-phone*
- *iPhone* “is a semantic sibling of” the *Samsung Android phone* (c.f. [Brunzel, 2008; Brunzel and Spiliopoulou, 2007; Rachakonda et al., 2012]).
- *iPhone* “contains” the *iOS* operating system

In the above example, the actor has embedded her individual world-view into her review in terms of various semantic associations of the concept *iPhone* with other concepts such as *smart-phone*, *iOS*, etc.

There also exist social interactions between chunks of content in both, repository spaces and social spaces. This essentially translates to the existence of social interactions between the corresponding CPs in these spaces. In this sense, information spaces are, in general, “socio-cognitive spaces”.

In the example above, we have discussed how the individual world-view of an actor gets embedded inside her review of the Apple iPhone 4S. Now, other actors could write their own reviews about the iPhone. Also, it is possible for the reviews generated by these various actors to socially interact with one another. These interactions could be in the form of hyperlinks to each other’s reviews (in the case of a web page) or comments to each other’s reviews (in the case of a blog post), or even further comments to each other’s comments (in the case of a blog post). Since postings, comments and web page content are manifestations of CPs, the interactions between them can be seen as interactions between the corresponding CPs.

In this fashion, when various CPs in an information space interact with each other, the individual world-views of different actors get aggregated in an emergent manner to form collectively held world-views of the multiple actors. We define such a process, in which individual CPs interact with each other to emergently give rise to commonly held world-views, as a *socio-cognitive process* (SCP). Essentially, the commonly held world-views of a population define the commonsense meaning ascribed to a given concept, which is expressed in terms of its associations with other concepts in an information space.

In other words, the commonly held world-views represent the *semantics* embedded within the information space. In this sense, semantics emerge

from the aggregation of individual or “local” world-views of the actors in a population.

1.2 Repository Spaces vs Social Spaces

Repository spaces and social spaces differ in three aspects: (i) boundaries within which social interactions take place between cognitive processes, (ii) engagement of actors across the space, and (iii) the localized synchrony of social interactions. This is explained below.

1.2.1 Boundary of Social Interaction

A key difference between repository spaces and social spaces is the boundary of social interaction between chunks of content (and hence their corresponding CPs). Even though social interactions between CPs play a crucial role in both these spaces, the boundary of social interactions in repository spaces is practically the entire repository, while the boundary of social interaction in social spaces is a single SCP and not the entire social space.

Example for Repository Spaces On the Web, in order for an actor Y to debate or argue about the ideas expressed by another actor X in a web page A , either (i) a new web page B must be created by Y containing a link to A together with Y 's own ideas, or (ii) an existing web page C owned by Y must be edited to include a link to A together with Y 's own ideas. Thus, it is possible for any document to link to any other document in a repository environment. The individual CPs of X and Y do not seem to be contained

within a well-defined boundary that allows multiple world-views to emerge coherently in a focused manner.

Example for Social Spaces In a social space such as Wikipedia, in order for an actor Y to debate about the ideas expressed by another actor X in a web page A , the actor Y need not use a separate document. Here, Y need only edit the document A to reflect his argumentations. Similarly, several other users may edit A to reflect their own ideas about the topic of A . Here, contrary to repository spaces, the individual CPs of the various actors are contained within a well-defined boundary (which is the SCP), which allows multiple world-views to emerge coherently within the boundaries of the SCP.

1.2.2 Scope of Actor Engagement

Another difference between repository spaces and social spaces is the scope of actor engagement, even though content is created in both these spaces by autonomous actors executing CPs. While actor engagement is highly focused towards a small set of documents (i.e. CPs) in repository spaces, actor engagement in social spaces can be spread across a large number of SCPs. In a repository space, an actor can typically edit only documents created by herself. In contrast, an actor in a social space can typically contribute to any existing SCP.

Example for Repository Spaces If actors X and Y own web pages A and B respectively, then X is allowed to edit only A , and not B . Conversely, a document on the Web can be edited by only a pre-determined number of

actors. In other words, web page A can be edited only by X (who is known to be the owner A), and not by Y . This type of isolated engagement results in social interactions that are limited only by the boundaries of the entire repository space.

Example for Social Spaces A user is allowed to edit any article in Wikipedia, and is not pre-determinedly restricted to accessing only a subset of Wikipedia articles. Conversely, an SCP can involve any number of actors. An example of this is multiple actors simultaneously editing a single Wikipedia document until an information architecture acceptable to most of the actors is evolved.

1.2.3 Localized Synchrony of Social Interactions

Even though social interactions take place within both, repository spaces and social spaces, they differ in terms of synchrony. In a repository space, argumentations and cogitations take place through the use of links and citations. However, since the entire repository space represents a single SCP, it is difficult to track documents pertaining to a given topic, so that further argumentations can be made about it by autonomous actors creating their own documents that cite existing documents. Thus, social interactions regarding a given topic in a repository space are typically spread over time in an asynchronous fashion, owing to their non-localized nature. On the other hand, in a social space, the thread of discussion on a given topic is limited to one or more localized SCPs, which the user is participating in. Thus, it is easy to track SCPs pertaining to a given topic, so that further argumentations and

cogitations can be made about it. Due to this, the social interactions in an SCP in a social space are typically synchronous. In other words, CPs are executed inside such an SCP within short durations of each other, giving the impression that a focused discussion is taking place within the SCP and that the CPs are “in tune with each other”.

Example for Repository Spaces Consider a research paper A being published in a digital library on a given topic. Assume that research paper A is a critique of an existing research paper B . Since there was no specific, well-defined boundary (or “container”) relating to B within which the critique presented by A could be posted, A was created as a separate document in the repository. Here, the repository itself is the “container” – the boundary within which an article should be posted is defined only by the boundaries of the repository. In other words, the social interactions between A and B would have to span the entire repository, and are therefore not localized. Now, it becomes difficult for the author of B to follow and to respond to A ’s arguments (and vice-versa, as time progresses). The scope of the discussion between A and B is unfocused, as the discussion takes place in “silos” (i.e. through the creation of more and more new documents) inside the repository space. This causes the social interaction in a repository space to spread over time and also become asynchronous.

Example for Social Spaces Consider a blog post on a given topic. For an actor interested in this topic, it is straight-forward to keep track of the developments in the discussions related to that post. This is because the

developments relating to that post take place through the use of comments, backtracks and votes in the context of the post itself. In other words, social interactions relating to this topic would then be localized to the blog post itself, and not the entire blogosphere. This makes it easy to follow the topic, and provide one's own world-views on that topic in short periods of time. Hence, comments and backtracks on the blog post are usually synchronous, and discussions appear to be focused.

Based on the above observations, we can draw the following interpretation about repository spaces and social spaces: *While a social space is characterized by the existence of several focused, synchronous SCPs, a repository space can be seen as representing a single broad-scoped, asynchronous SCP in its entirety.*

Figure 1.1 summarizes the difference between repository spaces and social spaces in terms of the nature of cognitive process and their social interactions.

It may be noted that, in addition to the above differences, repository spaces and social spaces can also be distinguished in terms of *a priori* well-understood ontological associations assumed by the population. In an SCP within a social space, the population typically holds some assumed ontological associations of some concepts. For instance, in a forum-thread on a technology forum like Slashdot, the population implicitly assumes that the term "Apple" refers to the consumer electronics company and not the fruit. In contrast, the population in a repository space tends not to have such *a priori* assumptions about ontological associations of concepts in their CPs.

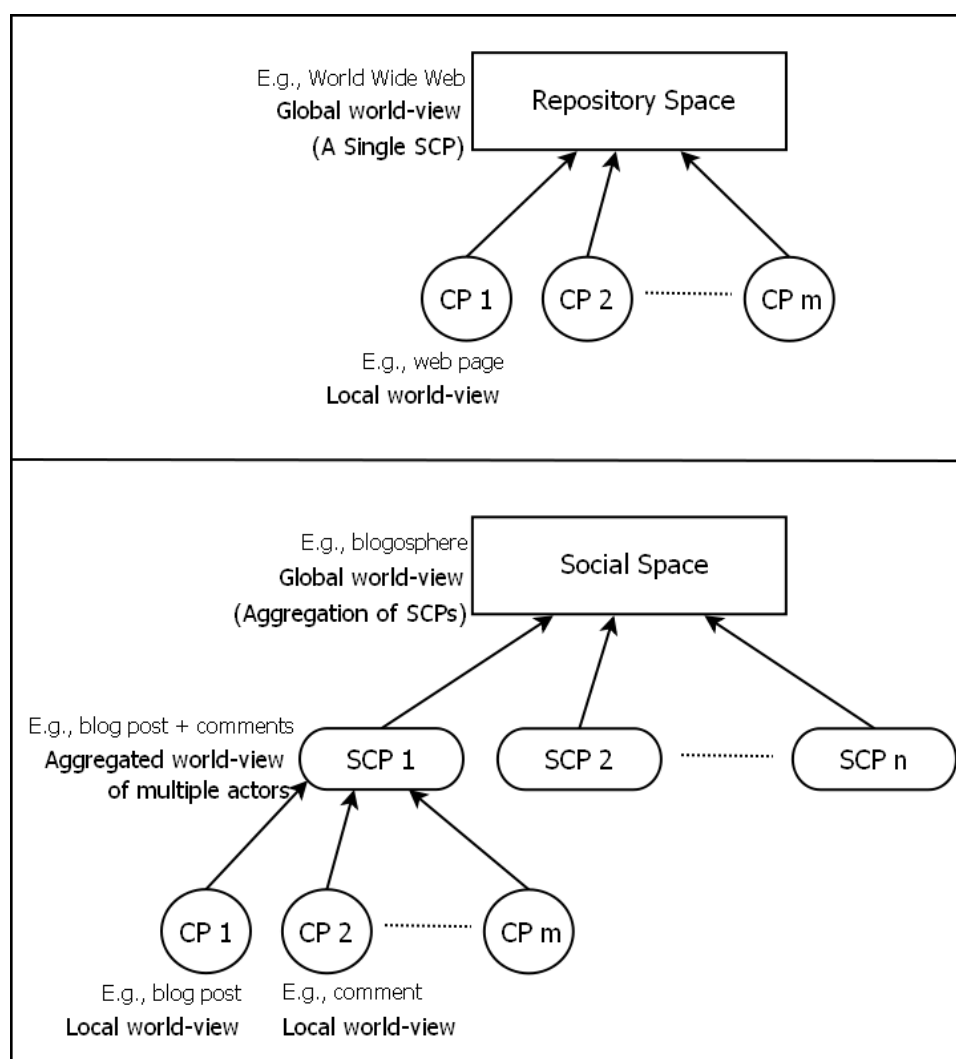


Figure 1.1: Illustration of the differences between repository spaces and social spaces

1.3 Semantics in Information Spaces

We now formally define repository spaces and social spaces.² These formal models also explain how semantics are embedded in information spaces as

²The work on formal modeling of social spaces was partly done in collaboration with [Mani, 2011].

world-views of actors.

1.3.1 Repository Spaces

A repository space is formally defined as:

$$S_R = (A_R, p_R, C_R, W_R) \quad (1.1)$$

where (i) A_R is the set of actors in the repository space, (ii) p_R is a single SCP capturing the social interactions of the entire space, (iii) C_R is a set of cognitive contexts (representing individual cognitive processes) that define cognitive activities such as document creation/editing, and (iv) W_R is the shared or “global” world-view of the actors across the space.³

Each actor in a repository space basically represents a local or individual world-view, and is modeled as:

$$\forall a \in A_R, a = (V_R, \varepsilon_a, R) \quad (1.2)$$

where (i) V_R is a set of concepts such as documents, topics and named entities within the repository space, (ii) R is a set of ontological *relationship types* (e.g., *is-relevant-to*, *is-important-in-the-context-of*, *is-an-endorsement-of*, etc.), while $\varepsilon_a \subseteq V_R \times R \times V_R$ is a set of *semantic associations* defining the local world-view.

Based on this, the global world-view W_R is defined as a semantic network (or ontology) in the form of a labeled multi-graph:

³It may be noted that multiple global world-views could emerge in an information space in isolation on a given topic.

$$W_R = (V_R, \varepsilon, R) \quad (1.3)$$

where $\varepsilon \subseteq V_R \times R \times V_R$ is the “global” set of semantic associations shared by the population of actors in general. While the global set of semantic associations held by the population is an aggregation of the semantic associations that are individually held by all the actors, it is not always a *union* of the individual semantic associations. For example, actor a could hold a semantic association such as “the Samsung phone is the best in the class of smart-phones”. However, not every semantic association held by every actor is *globally* held by the population at large. For example, the global world-view held by the population need not reflect actor a ’s world-view that the Samsung phone is the best in the class of smart-phones. We assert that the aggregation of the individual semantic associations is an *emergent process* in both, repository spaces and social spaces, resulting in the global set of semantic associations held by the population.

For a repository space, a cognitive context $c \in C_R$ is defined as a set of concepts that have occurred together within a given boundary, $c \in 2^{V_R}$. For the purposes of analytics, the boundary of a cognitive context can be defined in terms of a paragraph, a passage, an entire document, etc.

The socio-cognitive process p_R representing the social interactions of the entire space is represented as a sequence of cognitive contexts, $p_R \in C_R^*$. For instance: (i) the Web represents an SCP (i.e. a repository), where each document represents a cognitive context and links between documents represent social interactions between cognitive contexts; here, the actors are

the individual authors of web pages, (ii) a digital library such as CiteSeer⁴ represents a single SCP, with each scholarly paper as a cognitive context interacting with other papers through citations; here, the authors of the papers are the actors in the space.

In a repository space, while social interactions between cognitive processes may lead to the emergence of commonly held “world-views” of the population at large, the individual cognitive processes themselves do not arise from any form of social consensus. These CPs are executed by a single actor,⁵ and represent the local world-view of the individual actor. However, the social interactions between the CPs do result in emergent global world-views within the repository space. For instance, a Web search engine might be interested in knowing web pages or magazine articles that the population seems to consider most relevant to the topic of “Apple iPhone 4S”. Similarly, a focused crawler would benefit from finding out which hyperlinks from its current web page are considered most relevant to a given topic.

1.3.2 Social Spaces

A social space is formally defined as:

$$S_S = (A_S, P_S, C_S, W_S) \quad (1.4)$$

where (i) A_S is the set of actors in the social space, (ii) P_S is a *set* of socio-cognitive processes, (iii) C_S is a set of “socio-cognitive contexts” defining a

⁴<http://citeseer.ist.psu.edu/>

⁵Or by a group of actors who are in agreement with each other beforehand about the information architecture of the CP, and can therefore be viewed as a single entity

topical sub-unit in a socio-cognitive activity, and (iv) W_S is the shared or “global” world-view of the actors.

Each actor in a social space basically represents a “local” or individual world-view, and is modeled as:

$$\forall a \in A_S, a = (V_S, \varepsilon_a, R) \quad (1.5)$$

where V_S is a set of concepts and R is a set of ontological relationship types (e.g., *is-a*, *is-in*, *is-an-attribute-of*, etc.), while $\varepsilon_a \subseteq V_S \times R \times V_S$ is a set of semantic associations defining the local world-view.

Based on this, the global world-view W_S is defined as a semantic network (or ontology) in the form of a labeled multi-graph:

$$W_S = (V_S, \varepsilon, R) \quad (1.6)$$

where $\varepsilon \subseteq V_S \times R \times V_S$ is the global set of semantic associations shared by the population of actors in general.

A socio-cognitive context $c \in C_S$ is defined as a set of concepts that have occurred together within a given boundary, $c \in 2^{V_S}$. For the purposes of analytics, the boundary of a socio-cognitive context can be defined variously in terms of a paragraph, an article-section, a comment, a status message, etc., depending upon the underlying social space.

A socio-cognitive process $p \in P_S$ is simply represented as a sequence of socio-cognitive contexts, $p \in C_S^*$. For instance: (i) a wiki page represents an SCP, where each paragraph represents a socio-cognitive context; here, the actors are the various contributors who collaboratively edit content, (ii) a blog

post along with its comments represents an SCP where each paragraph in the post and each comment note represent a socio-cognitive context; here, the blogger herself and the various commenters are the actors. Socio-cognitive contexts are so called because, contrary to a repository space, the social interaction between cognitive contexts in a social space take place within the *well-defined* social boundaries of an SCP.

Moreover, these social interactions are synchronous. For instance, the comments written in response to a blog post are not only contained within the boundary of its own SCP, but also typically occur within short time-spans of each other. Thus, a single blog post, together with its comments, represents an independent SCP containing the aggregated world-views of the actors who participated in it. The blog post along with its comments can be collectively seen to have given rise to an aggregated information architecture evolved through social consensus between various actors (i.e. authors and commenters). SCPs contain semantics latent within them. For instance, a blog post on the topic of “Apple iPhone 4S” could have elicited several comments containing positive and negative reviews of the product. Therefore, a large set of such SCPs on the topic of “Apple iPhone 4S” could contain a “global” world-view of the population at large about the product (which the population itself may not be conscious of).

1.4 Mapping the Problem Domain

In this chapter, we have identified two classes of information spaces, namely repository spaces and social spaces. As described earlier, there exist sev-

eral differences between these two classes of information spaces. However, a commonality between these two spaces is that content is generated by the cognitive processes of humans. These cognitive processes essentially give rise to a series of cognitive contexts, which contain *concepts* such as named entities and topics. Also, there exist *social interactions* between cognitive contexts in both these spaces (e.g., comment replies to blog posts, citations to other documents, etc.).

In other words, there are two kinds of “artifacts” that both, repository spaces and social spaces, are associated with: concepts and social interactions. Semantics could be mined using either or both of these artifacts in an information space. The problem domain can therefore be mapped as shown in Table 1.1. Here, the columns correspond to the artifacts, while the rows correspond to the information spaces.

In this thesis, we address only a subset of this problem domain. In Chapter 4, we address the problem of semantics extraction in repository spaces such as scientific literature corpora and the Web, where the artifacts of interest to us are *citations* between documents, which define the social interaction between the cognitive contexts defining the documents.⁶ In Chapter 5, we address the problem of semantics extraction in social spaces such as Wikipedia, where the artifacts of interest to us are *concepts* defined by named entities (or noun phrases) contained in the cognitive contexts of that space.

We now briefly introduce the specific pain points that we address in this thesis, in terms of semantics extraction in repository spaces and social spaces.

⁶Here, we collectively use the term *citation* to mean hyperlinks in Web-based information systems as well as citations in scientific literature.

	Concepts	Social Interactions
Repository Spaces	E.g., (i) Building topic maps of scientific literature on Arxiv, (ii) Finding related documents on the Web, etc.	E.g., (i) Finding important articles in CiteSeer using citations, (ii) Finding important citations themselves on the Web, etc.
Social Spaces	E.g., (i) Finding the semantic attributes of a named entity in Wikipedia, (ii) Finding semantic siblings of a term in Twitter, etc.	E.g., (i) Finding important conversations on Facebook or Gmail using comments, replies and “likes”, (ii) Extraction of hierarchical structures based on trust/distrust patterns between actors in a signed network like Slashdot Zoo, etc.

Table 1.1: Semantics extraction in information spaces: The problem domain

1.4.1 Pain Points

Endorsed Citations in Repository Spaces

Consider a topical surfer in a repository space such as the ACM Digital Library (DL), who is interested in the topic of Signed Networks. Suppose she begins surfing from the ACM DL page on the paper *Signed networks in social media*,⁷ and hopes to continue surfing on this topic by following outgoing citations from this paper. However, the outgoing citations from this paper

⁷<http://dl.acm.org/citation.cfm?id=1753532>

are not *uniformly* relevant to the topic of Signed Networks. For instance, this paper cites (among others): (i) a paper entitled *Social capital in the creation of human capital*, which has its roots in sociology and economics, and (ii) a paper entitled *Structure balance: A generalization of Heiders theory*, which has its roots in Graph Theory. While the former is definitely relevant to the topic of Signed Networks, it can be argued that the latter is more relevant to this topic. Our topical surfer is likely to follow outgoing citations based on the topical relevance of the target document to the source document. In order to differentiate the topical relevance of outgoing citations from a given document, we rely on the co-citation patterns of the source and target documents of the given citation. We term such citations, which are topically relevant, as “endorsed” citations. We also quantify the degree of endorsement of a citation in terms of an endorsement probability.

Essentially, we contend that not all social interactions within a repository space are topically relevant or uniformly important. Endorsed citations can be used to guide topically focused crawlers too. We also use the idea of citation endorsement for topically ranking documents in a repository space. In literature, there do exist approaches for topically ranking documents (c.f. [Haveliwala, 2003; Kleinberg, 1999; Lempel and Moran, 2001]). However, these approaches do not distinguish between outgoing citations from a document, considering them to be uniformly relevant. Also, traditionally, topical ranking approaches have considered documents as artifacts of interest, with a focus on finding “important” documents using the citation structure (c.f. [Abiteboul et al., 2003; Page et al., 1999]). In contrast to this, we focus on mining important *citations* themselves [Mutalikdesai and Srinivasa, 2010],

and then using them to track documents of interest.

Object-Attribute Relationships in Social Spaces

Consider a product strategist trying to identify what concepts the population at large uses in describing the product *iPad* in a social space like Twitter. Assume that such an analyst would be interested in identifying a set of terms or “attributes” that, according to the population, collectively characterize the product *uniquely*. The analyst could use such information for Internet marketing or Adwords-like online advertising placement.⁸ Given an “object” like *iPad*, we define as its attributes those terms that collectively help in describing it uniquely, and each of which describes at least one of its properties.

It may be noted that not all terms that are topically relevant to the object can be its attributes. For instance, the terms *iPhone*, *iPad Mini*, *Galaxy Tab*, etc. are topically relevant to *iPad*. However, they do not describe *iPad* uniquely, as the same set of terms is relevant to *Aakash tablet* as well. Hence, these terms do not constitute the attributes of *iPad*. Therefore, topic modeling approaches such as Latent Dirichlet Association (LDA) [Anthes, 2010; Blei and Lafferty, 2007; Blei et al., 2003; Hofmann, 1999a] are not ideally suited for mining the attributes of an object in a social space.

In order to mine concepts that are deemed by the population at large to be semantic attributes of a given object, we rely on exploiting the co-occurrence patterns of concepts within cognitive contexts inside a social space. In addition to the above example, mining object-attribute relationships

⁸<http://adwords.google.com/>

finds applications in various forms of analytics and decision support, such as: (i) a market research analyst attempting to classify various Android phones based on the terms used to describe them by the population, and (ii) an academic researcher surfing on social bookmarking systems like delicious.com, trying to find articles “similar” to a given article in terms of their attributes.

In this thesis, we model the problem of mining semantics in information spaces in terms of mining the “global” or shared world-view of the population, which are embedded within an SCP. In order to mine such semantics, we use *co-occurrence analysis*, which is described in detail in Chapter 3. We posit that the co-occurrence patterns of concepts not only allow us to identify the meanings associated with the concepts, but also to identify the “type” of meaning they carry with respect to each other. In other words, co-occurrence analysis not only helps in the identification of latent semantics in information spaces, but also in identifying *semantic labels*.

1.5 Contributions of this Thesis

The contributions of this thesis can be summarized as follows:

- Modeling semantics extraction in information spaces
 - Distinction between repository spaces and social spaces as two classes of information spaces
 - Formal models for repository spaces and social spaces
 - Casting the problem of semantics extraction in information spaces as mining the global world-views of the population

- Motivation for the use of co-occurrence analysis for mining the world-views of the population
- Co-occurrence based semantics mining from the linkage structure in repository spaces
 - Interpretations of co-citations (co-occurrences of citations)
 - Notion of “endorsed” or important citations based on co-citation structures
 - Quantification of citation endorsements
 - Document ranking model based on endorsed citation structure
 - Exploratory study of the properties of networks formed due to endorsed citations
- Co-occurrence based semantics mining in social spaces
 - Definition of object-attribute relationships
 - Co-occurrence-based hypotheses for detecting attributes, defined in terms of:
 - * “usability” of an attribute along with the object
 - * positive probabilistic dependence of an attribute on the occurrence of the object

1.6 Organization of this Thesis

The rest of this thesis is organized as follows.

- Chapter 2 provides a detailed survey of literature pertinent to the work done in this thesis, and also positions co-occurrence analysis as our methodology for mining semantics.
- Chapter 3 provides the philosophical underpinnings for using co-occurrence analysis for semantics extraction in information spaces.
- We then focus our attention on mining semantics in repository spaces. In Chapter 4, we present various interpretations about what the co-occurrence of links/citations could mean. We then describe in detail one of these interpretations – co-citations as citation endorsements. We also describe how the notion of citation endorsements can be used for ranking documents in a repository space.
- We then turn to the problem of mining semantics in social spaces. In Chapter 5, we define the problem of mining object-attribute associations in social spaces. We propose two hypotheses for mining object-attribute associations based on co-occurrence analysis.
- We provide concluding remarks on the problem of mining semantics in information spaces in Chapter 6.

2

Related Literature

In this thesis, our work can be seen as being composed of three main parts:

- Modeling of semantics as the world-views of a population (in repositories and social spaces), and the use of co-occurrence analysis models for extracting such semantics. (Chapters 1 and 3)
- The analysis of co-occurrence patterns of interactions (i.e. citations) in repository spaces, for the task of extracting semantics. In particular, we

address the problem of extracting endorsed interactions. (Chapter 4)

- The analysis of the co-occurrence patterns of concepts (i.e. named entities in text) in social spaces, for the purpose of extracting semantics. In particular, we address the problem of extracting object-attribute relationships. (Chapter 5)

In this chapter, we survey literature relating to the above aspects of this thesis. This literature survey is, therefore, structured as described below.

In Section 2.1, we survey literature pertaining to the first part of this thesis, and position our motivation to use co-occurrence analysis in this thesis. This is followed by a survey of research in semantics extraction and information retrieval using co-occurrence analysis, in Section 2.1.2. Here too, we position our work with respect to existing approaches. In Section 2.2, we survey literature pertaining to the second part of this thesis, and position our work on endorsed citations. Finally, in Section 2.3, we address literature pertaining to the third part of this thesis, and position our work on detecting object-attribute relationships in social spaces.

2.1 Models for Semantics Extraction

We have modeled the problem of mining semantics as the problem of mining *global world-views* of the population in an information space. We use *co-occurrence analysis* as the methodology for mining such semantics. Two concepts in an information space are said to have *co-occurred* if they occur together within one or more cognitive contexts.

We first describe the existing approaches for mining semantics, outlining their shortcomings. We then introduce the motivation for using co-occurrence analysis for mining semantics, in comparison with existing approaches.

2.1.1 Existing Approaches for Semantics Extraction

The existing approaches for semantics extraction can be broadly classified into the following:

1. Dimensionality Reduction
2. Machine Learning
3. Generative Models
4. Network Models

Dimensionality Reduction

In dimensionality reduction approaches, the documents in the underlying corpus are represented as vectors over the set of all terms in the corpus. The entire corpus can thus be viewed as a collection of d column vectors (or d points in an t -dimensional space) with the t terms as dimensions (i.e. as a $t \times d$ term-document matrix).

Dimensionality reduction techniques work on the premise that documents do not uniformly occupy the entire space offered by all the dimensions; instead, documents occupy dense subspaces of the t -dimensional space, and are spaced closely. Hence, correlated terms (i.e. correlated dimensions) are collapsed using dimensionality reduction techniques such as singular value

decomposition (SVD), thus bringing semantically related terms closer to one another. Examples of such techniques include Latent Semantics Indexing (LSI) and its variants [Deerwester et al., 1990; Song and Park, 2007; Steyvers and Griffiths, 2007].

[Deerwester et al., 1990; Hofmann, 1999a,b; Steyvers and Griffiths, 2007] proposed the technique of LSI. They argued that semantically related terms (e.g., {*car, automobile*} or {*pet, cat*}) tend to co-occur within documents, thus leading the documents to occupy only semantically meaningful subspaces of the document space rather than the entire document space. They applied singular value decomposition to the term-document matrix and proposed a model for querying a smaller multi-dimensional space. However, while dimensionality reduction techniques succeed in identifying terms or concepts that are semantically associated with one another, they do not seem to be able to label the semantic association itself.

In order to describe our choice of co-occurrence analysis as the tool of interest in semantics extraction, we draw insights from the fields of *Ordinary Language Philosophy* (OLP) [Wittgenstein, 1953] and *Cognitive Science* [Foundalis, 2006]. OLP argues that the commonsense meaning acquired by a term is dependent on its usage context. In other words, the co-occurrence patterns of terms defines the meaning they acquire. Complementary to this argument, the Hebbian Theory of cognitive perceptions states that the human brain forms associations between observed concepts based on their co-occurrences with other concepts across a large number of “episodes” or occurrence contexts. In Chapter 3, we present detailed arguments that co-occurrence analysis is a manifestation of OLP as well as the Hebbian Theory.

We also argue that co-occurrence analysis goes beyond establishing just the meaning acquired by a concept within its occurrence context. Co-occurrence analysis also models the *type of meaning* acquired by a concept in relation to other concepts in its occurrence context. That is, co-occurrence analysis also models the *labels* of semantic associations existing between concepts, which define the world-view of the population in that space.

Machine Learning

Machine learning approaches typically use techniques like supervised learning, unsupervised learning and association rule mining for mining semantics such as classes, clusters and term associations from the underlying corpora [Agrawal et al., 1993; Ciaramita et al., 2005; Harish et al., 2010; Pang et al., 2002; Sebastiani, 2002].

Unsupervised learning approaches (e.g., [Ciaramita et al., 2005]) are successful in identifying clusters of semantically associated concepts based on observing patterns within the data. Similarly, supervised learning approaches have been used to identify semantic associations such as sentiments and opinions [Mullen and Collier, 2004; Pang and Lee, 2008; Pang et al., 2002] based on patterns observed in a training dataset. Association rule mining [Agrawal et al., 1993] can be seen as a special case of mining semantic associations.

However, none of these learning approaches necessarily explain why the observed patterns result in the mined associations. In contrast, we use co-occurrence analysis to mine semantics in terms of the global world-views of the population in an information space. The observed co-occurrence patterns result in semantics, because those semantics are embedded in the information

space by way of cognitive activities of actors. The co-occurrence patterns essentially encode the manner in which various concepts and their semantic associations are used by actors across cognitive contexts.

Generative Models

Generative models such as Latent Dirichlet Association (LDA) are popularly used to address the problem of topic modeling in a large corpus [Anthes, 2010; Blei and Lafferty, 2007; Blei et al., 2003; Hofmann, 1999a]. These topics are not known in advance, and are hence postulated using a hidden structure, which is then learned using posterior probabilistic inference.

In essence, a topic is defined as a probability distribution over words. A topic model is seen as a generative model for a document collection, where documents are assumed to be generated according to some elementary language model such as a hidden markov model [Andrews and Vigliocco, 2010], a bi-gram language model [MacKay and Peto, 1995], etc.

Estimating these topic distributions can be seen as a form of semantic association mining. However, parameter estimation in generative models is based on iterative optimization, and may converge to local optima, thus causing the convergence to vary across different experiments for the same corpus. On the other hand, modeling the co-occurrence patterns of concepts within an information space is not an optimization problem, and hence does not suffer from problems like convergence to local optima.

Network Models

Network models represent a corpus as a graph, where the nodes represent terms or entities, and edges represent relationships between the entities. Several network models for document retrieval and speech disambiguators [Ceglowski et al., 2003; Dagan et al., 1999, 1994] do not associate semantics with these relationships.

Semantic networks, in particular, represent the underlying corpus as a graph, where the edges represent the *observed* semantic relationships between terms [Glöckner et al., 2006; Quillian, 1968; Shapiro, 2000]. The relationships are captured using shallow parsing techniques scanning sentences for grammatical syntax [Carreras and Màrquez, 2005; Coppola et al., 2008; Pradhan et al., 2003]. Over such networks, graph-theoretic analyses like centrality, reachability, graph clustering, etc. can be used to extract further semantic associations.

However, the associations discovered using grammatical parsing techniques are heavily dependent on the linguistic structure of the documents in the corpus. On the other hand, the co-occurrence of concepts within a cognitive context is independent of linguistic structures. Therefore, modeling co-occurrence patterns does not involve grammatical parsing.¹

We now survey relevant literature on the use of co-occurrence analysis itself as a methodology for semantics extraction.

¹It may be noted that we model the co-occurrences of *named entities* or *noun phrases* in a text corpus. Grammatical parsing may be required for identifying the named entities or noun phrases. However, we use off-the-shelf algorithms or libraries for this task. This is not a part of the actual problem addressed in this work. In this work, we look to mine *semantic associations* between the named entities or noun phrases.

2.1.2 Co-occurrence Analysis for Semantics Extraction

Co-occurrence analysis has been used in information retrieval and semantics extraction since long. In 1977, [van Rijsbergen, 1977] proposed to use co-occurrence data from text corpora to mitigate the term-independence assumption commonly employed in traditional IR systems. He argued that, in reality, terms are not independent of one another. He captured the extent of dependence of two terms (upon each other) in terms of their co-occurrences.

[McDonald and Ramscar, 2001] experimentally verified the distributional hypothesis. [Rohde et al., 2004] also used co-occurrence analysis to derive the meanings of words in text corpora. [Terra and Clarke, 2003] used word co-occurrence to analyze the similarity of words in text corpora.

[Rapp, 2002; Wettler and Rapp, 1993] proposed a stochastic model for predicting the mental association of words using their co-occurrences in large text corpora. [Lund and Burgess, 1996; Lund et al., 1995] used a co-occurrence-based high-dimensional semantic space to model semantic memory. They too observed strong associations between pairs of words whose co-occurrence vectors clustered semantically according to multidimensional scaling. The above works are consistent with the principle of Hebbian Learning [Hebb, 1949].

[Patel et al., 1998] investigated the modeling of semantic representations in languages using co-occurrence statistics from large text corpora. [Sahlgren, 2006] proposed a model for semantic similarity of words, called the word-space model, based on lexical co-occurrence. [Veling and van der Weerd, 1999] used term co-occurrence patterns for resolving the word-senses (poly-

semy) of users' queries in an information retrieval system. They argued that capturing the co-occurrence patterns of the terms helps in grouping together terms that have semantic coherence.

It is interesting to note that dimensionality reduction models like LSI also model co-occurrences of terms [Deerwester et al., 1990; Hofmann, 1999a,b; Steyvers and Griffiths, 2007]. However, LSI is primarily aimed at conducting full-text querying in a term-document space with a reduced number of dimensions. LSI does not address the problem of mining the labels of semantic associations between terms.

[Rachakonda and Srinivasa, 2009a,b] used lexical co-occurrence patterns to identify the “topical anchors” of a given context. They defined topical anchors as terms whose semantics represent the topicality of the entire context.

[Essen and Steinbiss, 1992] addressed the problem of parameter smoothing for maximum likelihood estimation in stochastic language models. They used a term co-occurrence matrix as the confusion matrix for smoothing the conditional term probabilities of the language model. [Dagan et al., 1999, 1994] addressed the problem of likelihood estimation of word combinations for natural language processing applications using word co-occurrences.

In addition to lexical co-occurrence, several other forms of co-occurrence have also been analyzed. Association rule mining [Agrawal et al., 1993] can be seen as a form of co-occurrence analysis, wherein the goal is to mine sets of items that co-occur frequently within a database of transactions. Structural motifs such as co-linking/co-citation and co-tagging can also be seen as forms of co-occurrence (of URLs/citations and tags, respectively) (c.f. [Jin et al., 2009; Mutalikdesai and Srinivasa, 2010; Small, 1973, 1993]).

Co-occurrence analysis has also found applications in the life sciences. [Maskery et al., 2006] have used co-occurrence analysis to discover novel patterns of pathology for breast cancer. Similarly, [de Ridder et al., 2007] used co-occurrence analysis to discover a significant number of cooperating gene mutations, which play a role in the development of cancerous tumors.

In our work, we argue that co-occurrence analysis has its philosophical basis in the Ordinary Language Philosophy and Hebbian Learning. In contrast to the above works, we also argue that co-occurrence analysis helps not only in identifying the meaning acquired by a concept within a given context, but also in identifying the label of its meaning (or association type) with respect to other concepts within that context.

2.2 Co-citation Analysis in Scientific Literature and the Web

As mentioned earlier, we are interested in the co-occurrence patterns of interactions between documents in a repository space. In particular, we look to study the co-occurrence patterns of interactions in order to mine “endorsed” interactions in the repository space. Interactions in repository spaces are typically established through the use of citations between documents. In Chapter 4, we explore the notion of endorsed citations, which are essentially citations that are “more important” than other citations that emanate from the same source. This stems from our position that not all outgoing citations from a document are uniformly important. We posit that interactions be-

tween documents in a repository space are non-uniformly important. Therefore, we look at analyzing co-citations, i.e. the co-occurrence of citations, between pairs of documents in order to mine endorsed citations.

Co-citation analysis has long been employed in traditional scientific literature to discover meaningful insights from a literature corpus, e.g. to discover clusters of related articles, journals and authors. In 1973, [Small, 1973] analyzed co-citations in Particle Physics literature. He found that a high degree of co-citation is a better indicator of topical relatedness than bibliographic coupling. [Small, 1993] also studied the changes in the structure of co-citation graphs of scientific literature over six years, thus making interpretations about the growth of a topic of study.

[White and Griffith, 1981] studied author co-citation graphs of Information Science literature over seven years. They were able to map: (1) identifiable clusters of authors with similar interests, (2) topical proximity of clusters to one another, (3) centrality of authors in each cluster, and (4) proximity between authors within and across clusters. [Saka and Igami, 2007] used co-citation analysis to generate a topic map of Modern Science and to analyze how various research areas are related to each other.

[White and McCain, 1998] presented a comprehensive domain analysis of the Information Science discipline using author co-citation. They presented several insights such as specialty structure of Information Science across 24 years, authors memberships in one or more specialties, changes in authors influence and eminence over various sub-periods, changes in the subject interests of authors, etc. [Cottrill et al., 1989] analyzed author co-citations in Innovation Research Traditions literature and identified cognitive relations

between the works of two specialties of this discipline, viz. diffusion of innovations and technology transfer.

[Zhao, 2005] analyzed author co-citations by considering the first five authors of a cited paper. This is in contrast to some of the traditional methods, where only the first author of a paper is considered in the co-citation graph. Zhao reported more coherent author clusters using five-author co-citation compared to only first-author co-citation. However, Zhaos approach represented fewer specialties in the field whose literature was being studied, in contrast to the traditional approach of analyzing first-author co-citation graphs.

Co-citations have also been analyzed in the context of the World Wide Web to discover pages with related content. [Dean and Henzinger, 1999] proposed that two pages are related if they are highly co-cited, an idea endorsed by [Davison, 2000] as well. [Hou and Zhang, 2003] also used co-citations to find semantically relevant pages. [Reddy and Kitsuregawa, 2001] used co-citations to discover Web communities.

Hyperlink-Induced Topic Search (HITS) [Kleinberg, 1999] utilized the bipartite structures at the core of Web communities to determine good hub pages and authority pages pertinent to a given query. These bipartite cores correspond to co-citations of authorities by hubs. [Jeh and Widom, 2002] used a broader structural context beyond co-citations to define a measure of similarity, known as SimRank, between two given nodes in a graph. They introduced a recursive intuition of similarity: *two nodes are similar if they are referenced by similar nodes*. The base case here is that nodes are maximally similar to themselves.

[Efron, 2004a,b] used co-citations to determine the political orientation of web pages, viz. left-wing or right-wing. Based on the likelihood that a given document is co-cited with documents of a known political orientation, he modeled the extent to which the given page has the same political orientation. [Thelwall and Wilkinson, 2004] used co-citations along with bibliographic couplings and direct citations to find similar websites within the UK academic Web.

[Vaughan et al., 2007] examined why websites are co-cited – with Canadian universities as the case in point – and discovered that two co-cited universities were academically related with high probability. [Vaughan, 2006] also used co-citations to show the linguistic and cultural differences in the way Canadian universities are perceived by the population. [Vaughan and You, 2006] hypothesized that the number of co-citations to a pair of business websites is a measure of the similarity between the two businesses. Since similar businesses are competitors, Vaughan and You argued that co-citation data can be used to map business competitive positions.

[Larson, 1996] used co-citations as a measure of relatedness of pages, and visualized clusters of related pages on various topics like Geophysics, Climate, Remote Sensing and Ecology using multi-dimensional scaling. [Pitkow and Pirolli, 1997] also used co-citations for clustering web pages.

[Moise, 2003] proposed the idea of “focused co-citations”. She argued that due to the presence of several web pages with no particular topical focus, just counting the number of co-citations between pairs of pages is not a good enough measure of relatedness. She proposed that given a page A , any other page B that is co-cited with A should contribute to the *topical*

focus of A proportionally to the joint probability of co-citation of A and B . In other words, $TopicalFocus(A) = \sum_B \frac{|A^I \cap B^I|}{|A^I \cup B^I|}$, where A^I is defined as the set of pages that cite A .

In comparison to existing literature on co-citation analysis in repository spaces, we look towards using co-citations as a distinguishing feature for citations from a given document. The “distinguished” citations form the topical “backbone” of the underlying repository space, which can then be used for focused resource discovery.

2.3 Detection of Object-Attribute Relationships

In Chapter 5, we address the problem of mining semantic relationships from social spaces by analyzing the co-occurrence patterns of named entities within the text. In particular, we look at the problem of mining object-attribute relationships. We now review literature relevant to this problem from the following perspectives:

1. Attribute Labeling
2. Ontology Learning and Building
3. Attribute Detection of Actors in Social Media
4. Commonsense Knowledge Acquisition
5. Relationship Mining

2.3.1 Attribute Labeling

There exists a body of work on extracting data records (e.g., product information) from web pages and labeling them with their attributes (e.g., name, description, price) (c.f. [Zhu et al., 2005, 2006]). We model the problem of attribute assignment differently from these approaches. First, we formally define the notion of social spaces, and look at the problem of attribute detection in terms of understanding the *shared world-view* of actors about the *concepts* (e.g., noun phrases) in the social space. In contrast to the attribute labeling task in the above works, the attributes that a population assigns to concepts in a social space are *emergent* semantics. Also, these attributes typically do not follow fixed templates and are highly unstructured.

2.3.2 Ontology Learning and Building

The extraction of semantic associations can be interpreted as ontology learning from social spaces. There have been several efforts towards ontology learning from text corpora (c.f. [Buitelaar et al., 2005]). Among these works, [Poesio and Almuhareb, 2008] particularly address the problem of learning concept descriptions from text, based on the attributes of the concepts. Since their work is based on text corpora, they use textual patterns and dependency parsing for identifying attributes. In contrast, we concentrate on building an ontology of object-attribute relationships using only the *co-occurrence* of concepts. The use of co-occurrence analysis makes our techniques generic enough for mining object-attribute relationships in various kinds of social space data – text-based (wikis, blogs, etc.) as well as metadata-based (tags, labels, etc.).

Recent efforts have also been focused on building ontologies derived from folksonomies. [Mika, 2007] presents a tripartite model of ontologies for semantic social networks, comprising actors, concepts and instances. He demonstrates the building of light-weight ontologies of concepts (tags) and social networks of actors using this model. In comparison, we propose a model for social spaces in terms of the socio-cognitive processes of actors, and address the problem of mining the shared world-view of the population in terms of object-attribute relationships. There also exist several approaches that address the problem of building tag hierarchies (i.e. taxonomies comprised of generalization-specialization relationships) from social tagging systems [Benz et al., 2010, 2011; Heymann and Garcia-Molina, 2006; Schmitz et al., 2006; Schwarzkopf et al., 2007]. In contrast to these approaches, we look to build ontologies consisting of object-attribute relationships rather than taxonomical relationships.

2.3.3 Attribute Detection of Actors in Social Media

In the realm of social spaces, some effort has recently been focused on detecting the attributes of social media *users*, where these users' attributes (e.g., ethnicity, gender) are latent within their tweets and status messages [Rao et al., 2011; Rao and Yarowsky, 2011]. As we have explained earlier in this thesis, these users (or actors) are a part of the definition of social spaces. In contrast to the approaches of [Rao et al., 2011] and [Rao and Yarowsky, 2011], rather than looking at the attributes of these actors themselves, we look at the "global" world-view held by the population of these actors about

various concepts within the social space, in terms of the attributes of the *concepts*.

2.3.4 Commonsense Knowledge Acquisition

Another body of work relevant to ours is commonsense knowledge acquisition (c.f. [Barbu, 2009; Blanco et al., 2012; Cao et al., 2008; Chklovski, 2003]). Commonsense knowledge includes predicates such as “a container can hold liquids” and “a laptop is a computer”. In our work, we look to mine predicates which particularly capture object-attribute relationships between concepts in social spaces. We view these object-attribute relationships as emergently reflecting how the population describes an object.

Mining such attributes can be seen as a commonsense knowledge acquisition activity. However, the commonsense knowledge obtained in this process pertains to the meaning attached to an object by the population in a social space. The association of an attribute with an object need not itself be commonsensical in nature. However, the attributes collectively lend an identity to the object, which, from the population’s perspective, can be deemed as being commonsensical.

2.3.5 Relationship Mining

Relationship mining is another area which is relevant to our work. Mining object-attribute relationships can be seen as a type of relationship mining activity. Association rule mining [Agrawal et al., 1993] has traditionally been seen as a special case of relationship mining, where the relationships to be

mined are the frequent associations between items. The popular Entity-Relationship model [Chen, 1976] for database modeling, commonly known as the ER-model, captures relationships between entities in real-world data.

[Pradhan et al., 2003] addressed the problem of annotating unstructured text sentences with simple role-based relationships such as “*who did what to whom*”, using shallow semantic parsing. [Carreras and Màrquez, 2005; Màrquez et al., 2008] and [Coppola et al., 2008] also addressed the problem of semantic relation mining from the syntactic parse tree of text. [Kasneji et al., 2009] proposed a method called MING for discovering “informative” subgraphs from entity-relationship graphs (such as social networks, domain-specific knowledge bases and ontologies), which indicate the relationships between an input set of entities. [Wang et al., 2010] addressed the problem of mining advisor-advisee relationships from academic collaboration/publication networks.

[Sundaresan and Yi, 2000] studied patterns of occurrences of related phrases in Web documents, and addressed the problem of identifying relations (such as *book-author* or *acronym-expansion*) between them. [Hattori et al., 2008] proposed that property inheritance from a concept to its hyponyms can be used for identifying hyponymy relationships on the Web.

In the life sciences domain, [Chang et al., 2009] addressed the problem of mining relationships among genes, diseases and drugs related to G-protein-coupled receptors. [Lange et al., 2005] proposed a data structure called Data Linkage Graphs for capturing relationships between entities such as enzymes, pathways, proteins and diseases. [Rinaldi et al., 2006] also address the problem of mining domain-specific relationships such as protein-protein interac-

tions and gene interactions latent within annotated document corpora.

Another topic relevant to our work is that of statistical relational learning [Getoor and Taskar, 2007]. This area of work deals with probabilistically modeling domains which have rich relational structures as well as inherent uncertainties. Tasks such as the collective classification of entities given their relationships with other entities, link prediction between related entities, and entity resolution for identifying equivalent entities are addressed in this body of work. In contrast to the above works, we speculate that there exist a specific kind of semantic associations called *object-attribute* relationships among a given set of concepts.

In the next chapter, we describe the philosophy behind using co-occurrence analysis for mining semantic associations from information spaces.

3

Co-occurrence based Semantics Mining

Philosophical foundations for understanding the semantics latent within ordinary language were first explored in a school of philosophy known as *Analytic Philosophy* [Preston, 2007]. Several schools of thought emerged from Analytic Philosophy. Of these, *Ordinary Language Philosophy* (OLP) [Wittgenstein, 1953] is particularly relevant to our work.

3.1 Ordinary Language Philosophy and Co-occurrence

The key idea behind OLP is: *meaning is usage*. Early ideas of Analytic Philosophy were based on the argument that terms in ordinary language have *pre-determined* commonsense interpretations that are well-understood, without reference to their usage contexts [Russell and Slater, 1986]. OLP is in contrast to this, arguing that the commonsense meaning acquired by a term is dependent on its usage context. Sometimes, polysemy resolution or word sense disambiguation happens by way of the term acquiring its meaning through its usage context. However, the argument of OLP goes beyond establishing word sense. Even seemingly “meaningless” terms acquire meaning depending on the way they are used within a context. This is demonstrated by the way the term “jabberwock” acquires a meaning in the following passage:

I love to go around the city in my jabberwock. My jabberwock has a seating capacity of five, and runs on diesel. It has a mileage of 15 kilometers per liter. It comes with cruise control too.

Here, we can see that the term “jabberwock” acquires the commonsense meaning, namely *car (vehicle)*. It acquires this meaning because of its usage with the other terms within the passage: *seating capacity, diesel, mileage, cruise control*, etc. In other words, the meaning acquired by the term “jabberwock” is dependent on its *co-occurrence* with other terms within the cognitive context. In this sense, co-occurrence analysis can be seen as a manifestation

of the principles of OLP.

However, OLP does not entirely reject the commonsense meanings that are attached to some terms without reference to their usage context. In the above example, the meaning of the term “car” is still well-understood commonsensically. However, the association of “jabberwock” with this meaning of “car” is dependent on the way “jabberwock” is used in the passage.

3.2 Hebbian Learning and Co-occurrence

Co-occurrence analysis of concepts in an information space also represents a fundamental theory in Cognitive Science, which describes one of the ways in which the human brain perceives associations between concepts. This is the theory of *Hebbian Learning* [Hebb, 1949]. The Hebbian Theory states that neural cells that are activated simultaneously tend to get connected to each other.

The Hebbian Learning principle captures the co-occurrence of concepts, as they are perceived by the brain over a large number of “episodes”, and forms associations between them [Lamberts and Goldstone, 2005]. Here, by an “episode”, we mean a short duration of time during which the brain perceives one or more concepts simultaneously. An episode may be thought of as a cognitive context.

For instance, when the term *Hiroshima* is uttered, we immediately tend to recall the term *Nagasaki* or the term *nuclear bomb*. This association was formed as a result of observing these concepts together over a large number of episodes. In other words, the co-occurrences of these concepts in

a large number of cognitive contexts caused them to be associated with one another. Much of the work on mining word associations and verifying the distributional hypothesis¹ is based on the Hebbian Theory (e.g., [McDonald and Ramsar, 2001; Terra and Clarke, 2003]).

This leads to the interpretation that co-occurrence analysis mimics one of the fundamental ways by which human cognition works [Foundalis, 2006]. At this point, it may be recalled that we had earlier argued that semantics come to be embedded in information spaces due to *cognitive activities* of human actors. This is the reason for our choice of co-occurrence analysis as the methodology for mining semantics from information spaces.

3.3 Beyond OLP and Hebbian Theory

In this thesis, we assert that co-occurrence analysis goes beyond establishing just the meaning acquired by a term within its occurrence context. Co-occurrence analysis also models the *type of meaning* acquired by a concept in relation to other concepts in its occurrence context. That is, co-occurrence analysis also models the *labels* of semantic associations existing between concepts, which define the *world-view* of the population in that space.

For instance, the *synonymy* relationship between two concepts – e.g., *feature film* and *movie* – may be modeled using a hypothesis based on their co-occurrence patterns as follows: if the concepts *feature film* and *movie* are synonyms, then (i) *feature film* and *movie* are used to *replace* each other within a context – i.e. there is a very low probability that the concepts

¹The Distributional Hypothesis states that words that occur together within the same contexts tend to have similar meanings.

feature film and *movie* co-occur within a context, and (ii) the contexts in which the concept *feature film* occurs are similar to the contexts in which the concept *movie* occurs – i.e. terms that *feature film* co-occurs with are similar to the terms that *film* co-occurs with.² Thus, due to its co-occurrence patterns, the concept *movie* not only acquires a *commonsense meaning* (due to its usage within certain contexts), but also acquires a *semantic label* (in this case, synonymy) for its meaning in relation to the concept *feature film*.

Therefore, in order to address the problem of mining semantics in terms of world-views of the population in information spaces, we use co-occurrence analysis. We have tested this assertion in terms of: (i) differentiating between “endorsed” (or relevant) and non-endorsed outgoing citations from documents in repository spaces, and (ii) identifying semantic attributes of concepts in social spaces.

This concludes the first part of this thesis, in which we have positioned co-occurrence analysis as the methodology of choice for extracting the global world-views of a population in an information space. In the next part, we turn our attention to the task of mining semantics from repository spaces using co-occurrence analysis. In particular, we focus on drawing interpretations about the co-occurrences of links/citations, and how these interpretations can be used to distinguish between the outgoing links/citations from a given document. This is explained in the next chapter.

²This hypothesis can be extended further with the assertion that the *attributes* of the concepts *feature film* and *movie* are similar. In this thesis, we address the detection of such attributes using co-occurrence based hypotheses.

4

Co-citation Analysis in Repositories

Citation analysis has been a crucial element of library and information sciences, which studies implicit semantics within clusters of documents, authors and journals. A citation from a document A to another document B can be seen as an endorsement of B by A [Chakrabarti, 2003; Garfield, 1971; Park and Thelwall, 2003; White and McCain, 1989].

Co-citation – i.e. the *co-occurrence of citations* within a document – has been used as a measure of topical relatedness among articles and authors in

scientific literature [Small, 1973, 1993; White and McCain, 1989]. Quoting [Budd, 1999], “If the citing author is asserting a knowledge claim in citing specific texts, then there is something that inheres in those cited texts that influences the citing author.” The co-citation of two documents by an author can thus be seen as those two documents “co-influencing” the author. Therefore, according to the theory of Hebbian Learning, we argue that if two documents are co-cited a large number of times (i.e. they “co-influence” a large number of citing authors), then there is some kind of relatedness between their topics, which causes them to be cited simultaneously and independently several times.

While the above notion of topical relatedness can be used to explain co-citations based on Hebbian Theory, we provide additional interpretations about the nature of co-citations.

4.1 Interpretations of Co-citations

Since citations in scientific literature are temporal in nature, there are no reciprocal citations. Also, the outgoing citations of a scholarly paper cannot change once it is published. However, on the Web, a page may not only change its existing outgoing citations, but may also become aware of one or more of its incoming citations, and thus add reciprocal outgoing citations. On wikis, changes in citation structure are even more likely since pages can be edited by any user. These differences in the underlying citation processes reflect on the respective co-citation structures.

In order to understand the nature of co-citations better, we propose dif-

ferent interpretations of what a co-citation means. We have developed three interpretations of a co-citation based on different assumptions about how it could have been formed. These are explained in Sections 4.1.1, 4.1.2 and 4.1.3.

4.1.1 Endorsements of Citations

Consider two documents A and B that are co-cited by a third document C . Initially, there existed a citation from A to B that was very relevant. An actor first discovered page A , and then discovered page B through the citation from A to B . The actor then created page C citing both A and B . In the context of the Web, this is known as the *copying model* [Kumar et al., 2000]. In other words, C discovered A , and then “copied” its citation to B . The copying model was proposed as a generative model to explain the growth of the Web.

However, we interpret the above co-citation of A and B by C as follows: C was able to discover B due to the citation from A to B ; in this sense, the co-citation by C can be seen as an *endorsement* of the citation from A to B . This is illustrated in Figure 4.1. Similarly, if a large number of users followed this citation, and ended up creating their own documents on a similar topic and citing both A and B , then we could view the endorsement of the citation as a global world-view of the population.

Such a global world-view can be used to distinguish “important” citations from the rest. Suppose A contained citations to a number of documents, but among those, if B alone has been highly co-cited with A , we can conclude

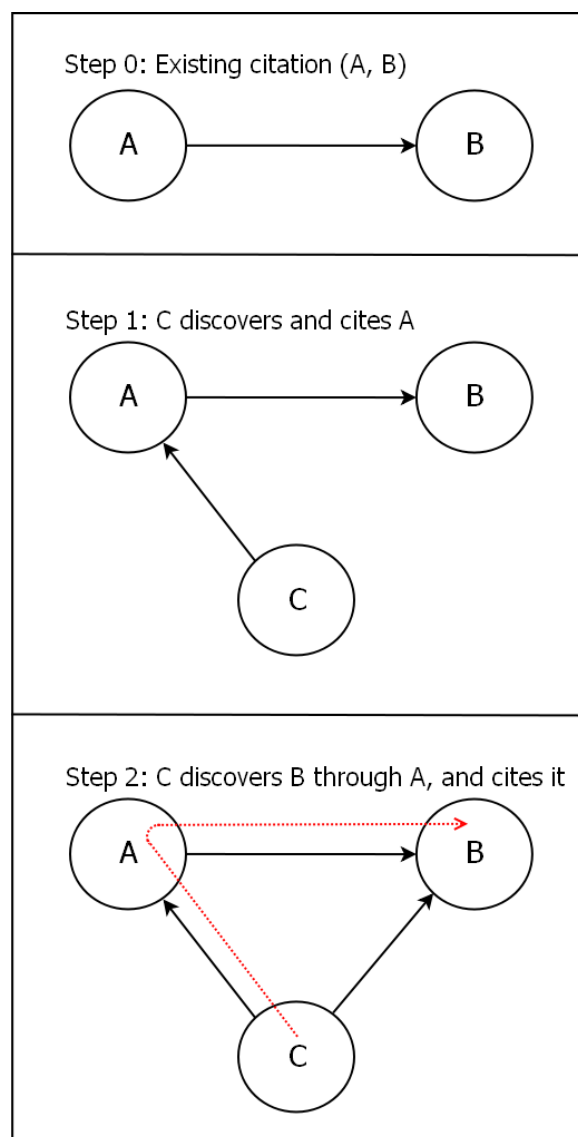


Figure 4.1: The process of a co-citation endorsing a citation

that the citation from A to B is more “important” than the rest of the citations by A [Mutalikdesai and Srinivasa, 2010].

In 1973, [Small, 1973] showed that, in scientific literature on Particle Physics, two articles that are highly co-cited are also directly connected

through a citation with high probability. A similar behavior is also known to have been observed in the case of web pages – two pages that are highly co-cited are also directly connected through a citation with high probability [Reddy et al., 2006]. However, on Wikipedia, the modal citation path length has been observed to be 2, given a pair of highly co-cited pages – i.e. the highly co-cited pages were not directly linked to one another; instead, with high probability, they were linked via an intermediary page [Reddy et al., 2006].

The above interpretation of co-citations, therefore, seems to be relevant to repository spaces such as the Web and digital libraries, and less so to social spaces such as Wikipedia. Since wiki pages are editable by anyone, it is more likely that the linkage structure keeps undergoing changes until a common information architecture emerges that is acceptable to most of the users involved. In web pages and scientific literature, however, citation happens independently – a citation created by one user cannot be edited by another. This enables us to view co-citations as citation endorsements in repositories and not in social spaces. We address this interpretation in detail in the rest of this chapter.

4.1.2 Knowledge Aggregation

Consider two documents A and B that are co-cited by a third document C . Documents A and B contain topically relevant content, irrespective of whether they cite each other or not. An actor discovered these documents, and created C , which cited both A and B . In other words, C acts as an

aggregator of the knowledge contained in A and B . In this sense, the co-citation of A and B by C can be seen as a *knowledge aggregation* activity by C . We can think of the co-citing document as representing a more general topic of A and B (e.g., survey papers, hub pages, directories). Figure 4.2 shows a co-citation interpreted as a knowledge aggregation activity.

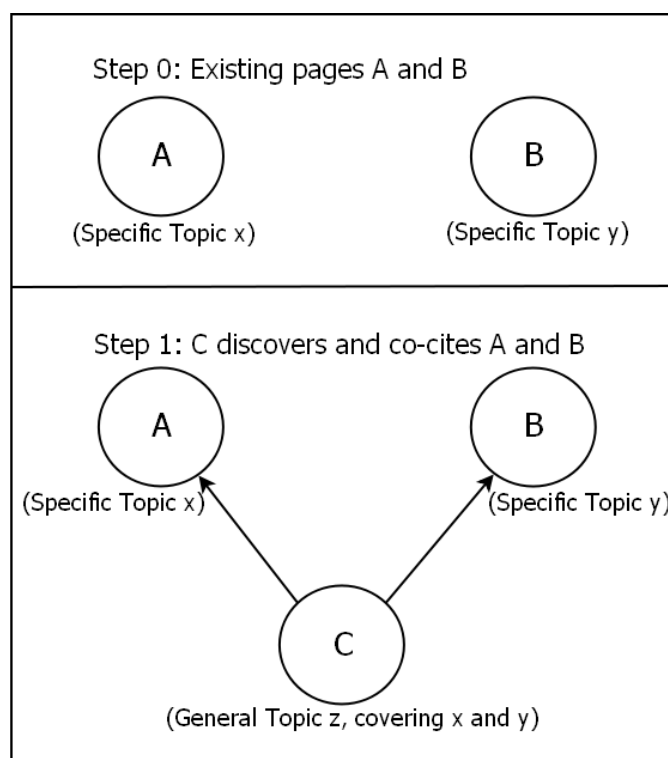


Figure 4.2: A co-citation as a knowledge aggregation activity

As discussed above, the co-citing document C can be seen to be on an aggregated topic (of A and B). As such, in the above example, if document A were to further co-cite two or more other documents, then the resulting “aggregation hierarchy” can be seen to represent a topic hierarchy.

Here, we assume that documents cite each other due to topical relevance.

However, this need not always be the case. In a repository space such as the Web, this interpretation of a co-citation pertains to the local world-view of a single actor.¹ In order to address this problem, the following heuristic can be employed: A document Y can be considered to be an aggregator of the topics of a set of documents \mathbf{X} , if the documents in \mathbf{X} have been co-cited by a large number of other documents as well. A large number of documents independently co-citing the documents in \mathbf{X} can be interpreted as being the global world-view that the documents in \mathbf{X} are related to one another.

Given such an aggregation hierarchy, if each individual document in this hierarchy can be labeled with a topic, then *topic classification* can be seen as one of the applications of co-citations as knowledge aggregation activities in repository spaces. However, we do not address this problem as part of this thesis, and view it as a potential line of future work.

4.1.3 Conditional Relevance

Consider two documents A and B . Suppose A is relevant to a topic t , while the topic of B is unknown. Now, if a document C cites A , we assume that C is also relevant to the topic t . In addition to this, if C cites B as well, then we interpret this co-citation of A and B by C as follows: *The world-view of C is that B is also relevant to the topic t .* If A , B and C were documents in a repository space, then this interpretation would be the local world-view of an actor in the cognitive context defining C .² Nonetheless, a global world-

¹However, in the case of a social space like Wikipedia, this interpretation reflects the aggregated world-view of multiple actors, contained within a single SCP. This is because, in a wiki, several users collaboratively edit a page to form co-citations.

²However, in a social space like Wikipedia, this interpretation would be the aggregated world-view of multiple actors within an SCP.

view does emerge from this interpretation in repository spaces, which can be described as follows.

A number of documents cite A , and among them, a few also cite B . The proportion of documents citing A , which also cite B , can be viewed as an indicator of the relevance of B to A . In other words, if we know that A is relevant to some topic t that defines the set of documents citing A , the co-citations can be seen as representing the conditional probability of topical relevance: *Given that A is relevant to t , what is the probability that B is relevant to t as well?* Figure 4.3 illustrates this interpretation. In other words, this interpretation gives us the global world-view of the “usability” of B as a reference, in a context that uses A as a reference.³

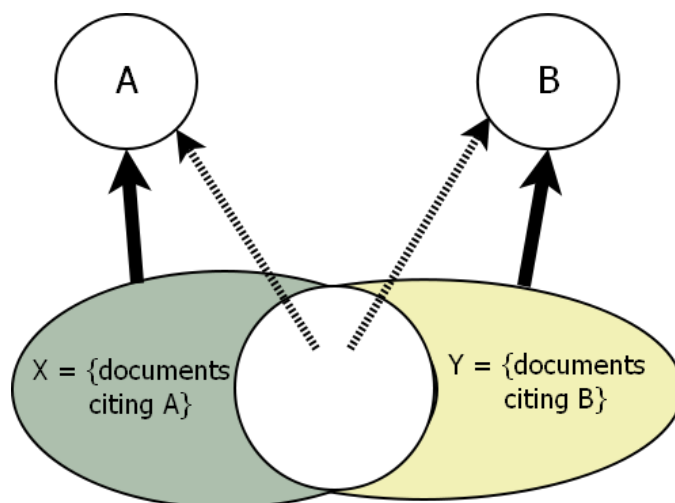


Figure 4.3: Co-citations as indicators of conditional relevance. The relevance of B to A can be given by $\frac{|X \cap Y|}{|X|}$.

In conjunction with the copying model [Kumar et al., 2000], the condi-

³In the next part of this thesis, we use this notion of “usability” or conditional relevance for modeling co-occurrences of concepts in a social space too.

tional relevance interpretation could also be used to predict the appearance of a citation between two documents in the future in a repository space. In literature, this is also known as the *link prediction* problem (c.f. [Al Hasan et al., 2006; Liben-Nowell and Kleinberg, 2007; Lü and Zhou, 2011; Sarkar et al., 2012; Song et al., 2009; Soundarajan and Hopcroft, 2012; Wang et al., 2011]). Let us assume that a document A cites another document B . Suppose \mathbf{C}_A is the set of documents citing A . Let $\mathbf{C}_{AB} \subseteq \mathbf{C}_A$ be the set of documents that cite B given that they also cite A . Now, the probability P_{CB} that a document $C \in \mathbf{C}_A$ “copies” A ’s citation to B , and itself ends up citing B , is given by $P_{CB} = \frac{|\mathbf{C}_{AB}|}{|\mathbf{C}_A|}$.⁴ However, we do not go into the details of this problem as part of this thesis, and view it as a part of future work.

In this work, we focus primarily on the first interpretation, *co-citations as citation endorsements*, which serves as the global world-view of the population about the distinction of certain citations from the rest in a repository space. We formalize the notion of citation endorsement as the probability with which any given citation from a given document points to another topically relevant document. Such a measure could be used to differentiate between citations from a document, based on their topical relevance to the document. It can hence be used for guiding surfers or crawlers seeking to maintain topical focus.

In Section 4.2, we describe a mathematical model for endorsed citations, and discuss our experiments on analyzing endorsed citations in a Web crawl as well as CiteSeer (a digital library).

⁴It may be noted that, in general, a given document C may copy the link to B from one of the various documents it cites – represented by the set \mathbf{A} – such that every $A \in \mathbf{A}$ cites B .

4.2 Co-citations as Citation Endorsements

Given a document C , we shall use the term C^O to refer to the set of all documents cited by C such that $C \notin C^O$. The *citation set* of C , denoted as C^I , is the set of all documents that cite C . In other words, $C^I = \{D | C \in D^O\}$. Given any set of documents $S = \{C_1, C_2, \dots, C_n\}$, the *co-citation set* is defined as:

$$CoCit(S) = \bigcap_k C_k^I \quad (4.1)$$

In this work, we shall be primarily concerned with co-citation patterns across *pairs* of documents. For notational simplicity, for any pair of documents $\{A, B\}$, we denote $A \cdot B = CoCit(\{A, B\}) = A^I \cap B^I$. We shall use the notation $A \Rightarrow B$ to denote a citation from A to B .

Given any pair of documents $\{A, B\}$ such that $B \in A^O$, the *endorsement probability* of the citation, denoted by $\rho(A \Rightarrow B)$, is computed as:

$$\rho(A \Rightarrow B) = \frac{|A \cdot B|}{\sum_{\forall X \in A^O} |A \cdot X|} \quad (4.2)$$

Consider Figure 4.4 depicting a fragment of a citation graph. Document A cites four documents, viz. B , C , D and E . Table 4.1 lists co-citation sets for A with each of its out-neighbors⁵ along with the corresponding citation-endorsement probabilities.

In our example above, we have not thresholded the co-citations. Co-cited pairs of pages on the Web are known to exhibit a power-law distribution in

⁵The out-neighbors of a document A are defined as the documents cited by A .

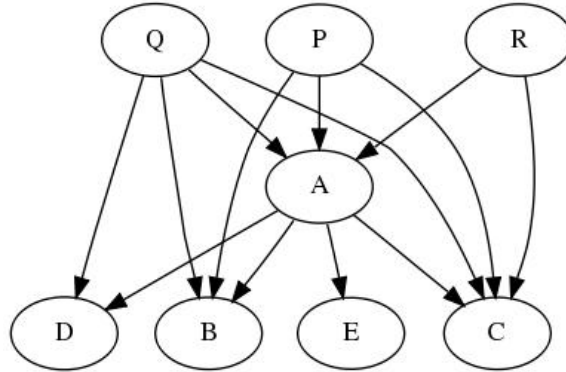


Figure 4.4: Sample citation graph

<i>Document Pair</i>	<i>Co-citation Set</i>	$\rho(A \Rightarrow \text{out-neighbor})$
$\{A, B\}$	$\{P, Q\}$	0.3333
$\{A, C\}$	$\{P, Q, R\}$	0.5
$\{A, D\}$	$\{Q\}$	0.1667
$\{A, E\}$	\emptyset	0

Table 4.1: Co-citations of document *A* with its out-neighbors and the corresponding citation-endorsement probabilities

the number of co-citing pages [Reddy et al., 2006]. We observe a power-law distribution in the number of co-citation counts⁶ in the CiteSeer dataset as well. Based on these distributions, we threshold the co-citations to include only those pairs of documents whose co-citation count equals at least a specified value. This way, we can be reasonably sure that two documents are genuinely relevant to one another, as they have been co-cited by a non-trivial number of other documents.

An alternative perspective for citation endorsements is as follows. Assume

⁶For a given pair of documents, the co-citation count is defined as the number of documents co-citing them.

a category of users who browse a scientific literature corpus for researching a given topic and end up creating a document of their own on that topic. The importance such users assign to other documents is based on the relevance of those documents to the topic of their interest. Given that one such user has cited a given document A , the endorsement probabilities of the citations from A can be seen as a relative measure of the tendency with which the user would cite any of A 's out-neighbors as well.

In the context of the Web copying model (c.f. [Kumar et al., 2000]), this can be seen as the probability of a citation being “copied” by a topically focused user. On a similar note, the citation-endorsement probabilities can be seen as the propensity that a topically focused crawler would index an outgoing citation relative to the other outgoing citations from its current document.

4.2.1 Endorsed Citation Graph

Given a citation graph $G = (V, E)$, where V is the set of documents and $E \subseteq \{(u, v) | (u, v \in V) \wedge (u \neq v)\}$ is the set of citations between documents, we can construct its *Endorsed Citation Graph* (ECG) as $G' = (V, E', \rho)$, where $E' \subseteq E$ is the set of ordered pairs of documents such that $(A, B) \in E'$ iff $B \in A^O$ and $|A \cdot B| \geq \delta$, where δ is a thresholding parameter defining a lower limit on the number of co-citations. The edges in E' are referred to as endorsed citations.

Here, $\rho : E' \rightarrow [0, 1]$ is a weight assigned to every $(A, B) \in E'$ indicating the endorsement probability $\rho(A \Rightarrow B)$ defined earlier in Equation 4.2.

The ECG provides us with a topical “backbone” inside the citation graph of a scientific literature corpus. Along similar lines, we can define the ECG for a Web crawl.

Handling Nepotistic Citation Endorsements in the Web

Arguably, a major source of co-citations in web pages are navigational hyperlinks, which are likely to be present across most of the pages within a website. In order to prevent this from affecting our interpretation, we first remove from consideration all nepotistic citations – i.e. citations originating within the same website – when counting co-citations for a given pair of pages.

In our analysis of the ECG for a Web crawl, we consider a citation between two pages belonging to the same parent host as non-nepotistic if they belong to different sub-domains. In other words, we treat *ab.xyz.com* and *wv.xyz.com* as two autonomous websites. Even though these websites are affiliated to the same parent domain, viz. *xyz.com*, we assume that they represent two self-contained sub-organizations. We assume that any citation from one of these websites to the other is not nepotistic. Likewise, a citation between two pages belonging to *ab.xyz.com* and *xyz.com* respectively is assumed to be non-nepotistic. While we realize that it is possible for a nepotistic citation to exist between two pages belonging to different websites, we do not concern ourselves with the problem of multi-host nepotism (c.f. [Chakrabarti, 2003]).

In some cases, navigational hyperlinks between websites need not necessarily have been created due to the “top-down” imposition of an information architecture. They may have independently *evolved* over time, in which case

the interpretation of co-citations as citation endorsements could still provide insights into the emergent information architecture of the Web.

It may be noted that, while there may also be several motivations other than navigation for the creation of citations between pages (c.f. [Thelwall, 2004]), those interpretations of the underlying citation do not affect our interpretation of a co-citation as a citation endorsement. Thus, for a given web page A , if $hname(A)$ is defined as its URL hostname, then:

$$A^I = \{C | ((A \in C^O) \wedge (hname(C) \neq hname(A)))\} \quad (4.3)$$

Note that if a co-citation $A \cdot B$ is under consideration, it could well be the case that A and B belong to the same website. However, this co-cited pair is considered non-trivial if A and B have a non-trivial number of co-citations coming from *other* websites (to which neither A nor B belong). Along similar lines, it can be argued that, in the case of digital libraries of scientific literature, we must not consider co-citations emanating from “self-citations” – i.e. citations between articles written by the same author(s) – while building the ECG. However, in our analysis of the ECG for CiteSeer, we assume that a citation from one article to another is based on merit, even if the two articles have common authors, since most scholarly articles are deemed to have undergone peer-review.

4.2.2 Experimental Analyses

We now describe the experimental analyses we performed over: (i) the ECG for a large Web crawl, and (ii) the ECG for CiteSeer.

Web Crawl Dataset We have performed our experiments on a Web crawl obtained from Ask.com⁷ in January 2006.⁸ It contains 10,623,000 pages and 85,812,128 citations in all. We removed all isolated pages from the crawl (i.e. pages having no incoming citations or outgoing citations). We also removed all self-citations (i.e. citations connecting a page back to itself). We were thus left with 8,430,736 pages and 84,460,523 citations. The indegree and outdegree distributions for this cleansed dataset follow a power-law with exponents of 1.88 and 3.42 respectively. We also observe that the initial segment of our outdegree distribution deviates from the power-law, thus suggesting that pages with a low outdegree might follow a different (Poisson-like) distribution. These observations are, in essence, consistent with those of [Broder et al., 2000].

CiteSeer Dataset We downloaded a snapshot of CiteSeer in February 2009.⁹ It contains 716,772 articles and 1,744,619 citations in all. These articles belong to the broad area of Computer and Information Sciences. Some of the citations were found to be pointing to other articles that were not included in the dataset. Hence, we removed such citations in order to obtain a self-contained snapshot of the CiteSeer citation graph, which contained 1,740,331 citations. The indegree and outdegree distributions for this cleansed dataset follow a power-law with exponents of 3.07 and 4.2

⁷<http://www.ask.com/>

⁸Thanks to Tao Yang and Ambuj K. Singh of UC Santa Barbara for providing this Web crawl.

⁹CiteSeer.IST scientific literature digital library (open archives initiative): The dublin core standard with additional metadata fields, including citation relationships (References and IsReferencedBy), author affiliations, and author addresses. http://cs1.ist.psu.edu/public/oai/oai_citeseer.tar.gz. Last accessed 6 February 2009.

respectively. It has previously been observed by [Redner, 1998] and [Bilke and Peterson, 2001] that citations follow a power-law indegree distribution in Physics literature as well. However, we observe that the initial segments of both, the indegree distribution and the outdegree distribution of CiteSeer differ significantly from the power-law, suggesting that articles with a low indegree/outdegree exhibit a different distribution.

We conducted all the experiments described in this chapter on a computer having a 4-core Intel Xeon processor with IA-64 architecture, 6 GB of RAM and a 200 GB hard disk drive. The algorithms for discovering ECGs and computing ERanks (explained later on in this chapter) were implemented in Java, with MySQL as the database for storing our datasets.

Analysis of the Web Crawl ECG

We found that 911,411 citations – i.e. 1.08% of all the citations in the Web crawl – have been endorsed by at least one non-nepotistic co-citation. As mentioned earlier, we define a citation as being nepotistic if its source page and target page have the same hostname. We also found that the number of pages upon which these endorsed citations are incident is 97,790 – i.e. 1.16% of the total number of pages in the crawl.

The non-nepotistic co-citations that endorse citations on the Web exhibit a power-law in the number of co-citation counts, as shown in Figure 4.5, with an exponent of 1.69. In this thesis, we use Pareto cumulative distributions (c.f. [Adamic; Adamic and Huberman, 2002]) to depict power-law distributions. This allows us to easily fit a linear regression to the distribution. The Pareto cumulative distribution is a power-law with an exponent of $\alpha - 1$,

where α is the power-law exponent of the original probability distribution.

Based on the distribution in Figure 4.5, we choose a co-citation count threshold of 10 for the Web crawl, since a very small number of citations have been endorsed by 10 or more non-nepotistic co-citations. This indicates that these citations might genuinely connect topically relevant pages.

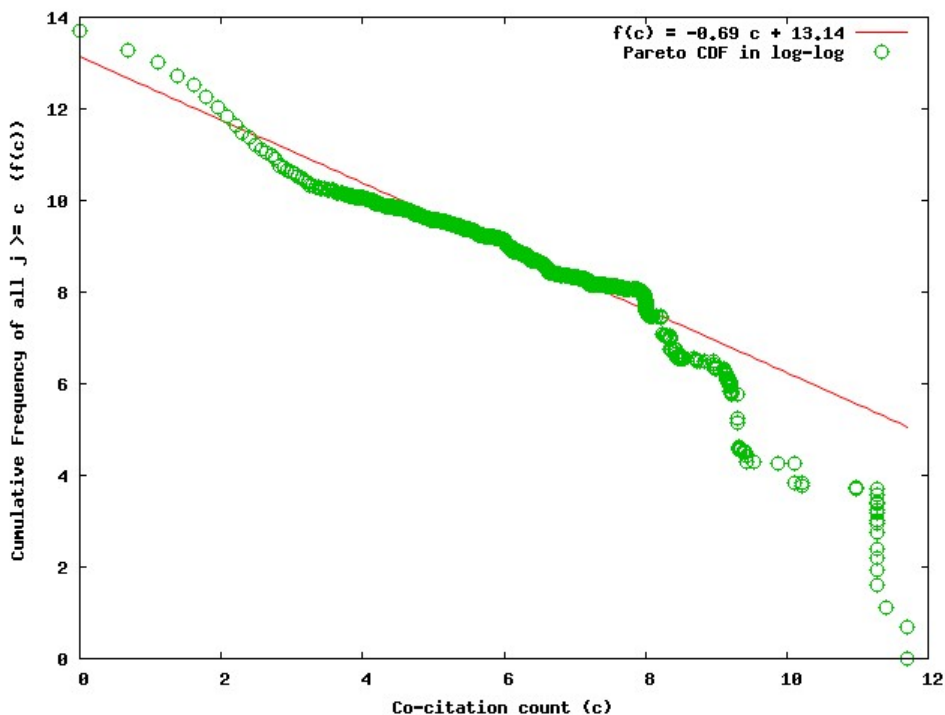


Figure 4.5: Pareto cumulative distribution (log-scale) of co-citation counts for endorsed citations in the Web crawl

We notice that only 97,858 citations in the Web crawl have been endorsed by 10 or more non-nepotistic co-citations. This forms only 10.74% of all the non-nepotistically endorsed citations and only 0.12% of the total number of citations in the crawl. These citations, along with the pages on which they are incident, form our ECG. The number of pages in this ECG is 11,180.

The endorsed citation data forms a very small percentage of the overall Web crawl. We believe that this is all the more reason to study semantics inherent in endorsed citations. Although citation endorsements may not affect crawling and ranking for generic searches in repository spaces such as the Web, they are much more relevant to topic-sensitive crawls and searches. Given that citation endorsements are rare on the Web, metrics based on citation endorsements can be seen as fine-tuning measures for relevance ranking.

In the Web crawl, we measured the proportion of outgoing citations from a given page that are endorsed. Even though the modal percentage of endorsed outgoing citations from the pages in our Web crawl ECG is 100, we found that, with a probability of 0.727, the pages have at most 60% of their outgoing citations endorsed. This statistic could be crucial to topical surfers, as a large majority of pages in the Web crawl have only a few outgoing citations that are deemed topically relevant to them.

The indegree and outdegree distributions for the ECG are shown in Figures 4.6 and 4.7 respectively. We see that the indegree distribution follows a power-law with an exponent of 2.6, but has an initial segment deviating from the power-law.

The outdegree distribution, on the other hand, seems to follow a Poisson distribution, and is therefore significantly different from the outdegree distribution for the Web in general.

As a result of the long-tailed indegree distribution, there are a large number of pages with only a few endorsed incoming citations, while there are only a few pages with a large number of endorsed incoming citations. The latter can be seen as topically authoritative pages to which several pages

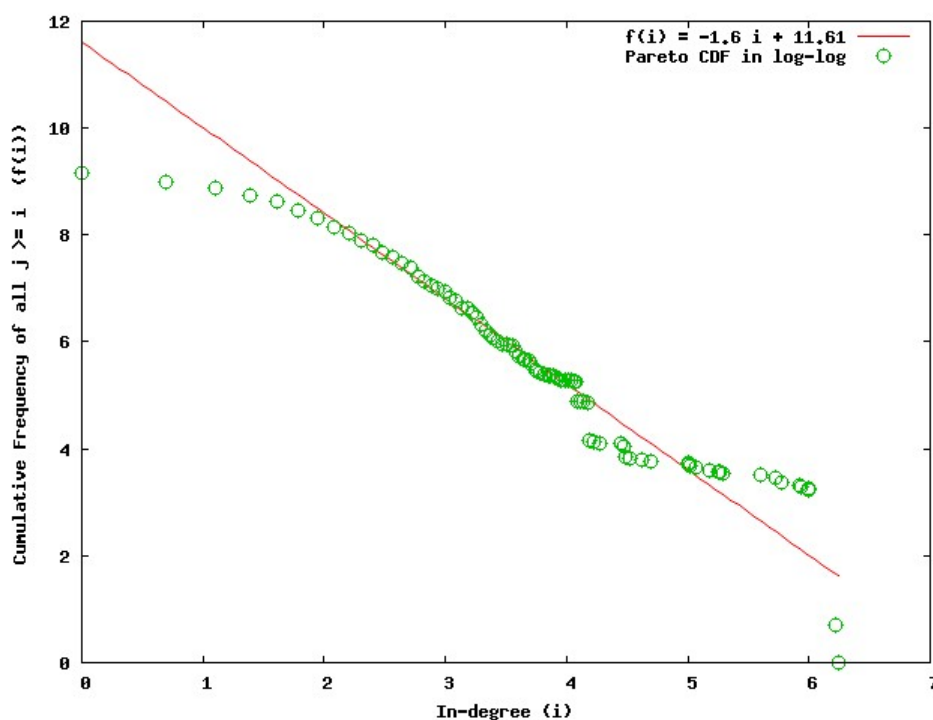


Figure 4.6: Pareto cumulative distribution (log-scale) of indegrees for the Web crawl ECG

have endorsed citations.

In the ECG, the number of pages with a non-zero indegree is comparable to the number of pages with a non-zero outdegree (85.02% and 95.55% respectively). However, while the maximum indegree here is 511, the maximum outdegree is only 68. This implies that even the best topical hubs in the ECG do not point to a very large number of pages. However, the top authorities are “recommended” very highly with large numbers of endorsed citations to them.

The ECG is a *disconnected graph* with 1,474 components. The distribution of component sizes is shown in Figure 4.8. With an exponent of 2.5,

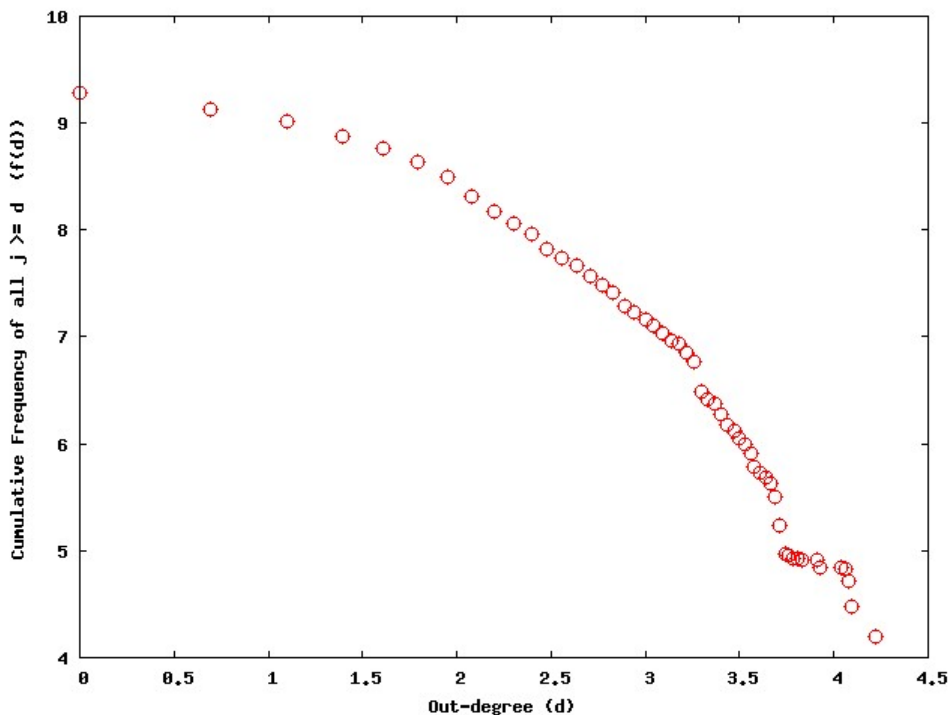


Figure 4.7: Pareto cumulative distribution (log-scale) of outdegrees for the Web crawl ECG

component sizes in the Web crawl ECG seem to follow a power-law – i.e. a small number of components contain a large number of pages, while a large number of components contain only a small number of pages.

We also observed that, in 649 of the 1,474 components (i.e. in 44.03% of the components) of the Web crawl ECG, pages belong to multiple websites. We found that 30,137 (i.e. 30.8%) of the citations in the ECG connect pages across different websites. This indicates that the notion of citation endorsement allows us to identify citations from a page, which point to other topically relevant pages, irrespective of the affiliations of the source and target pages of the citations.

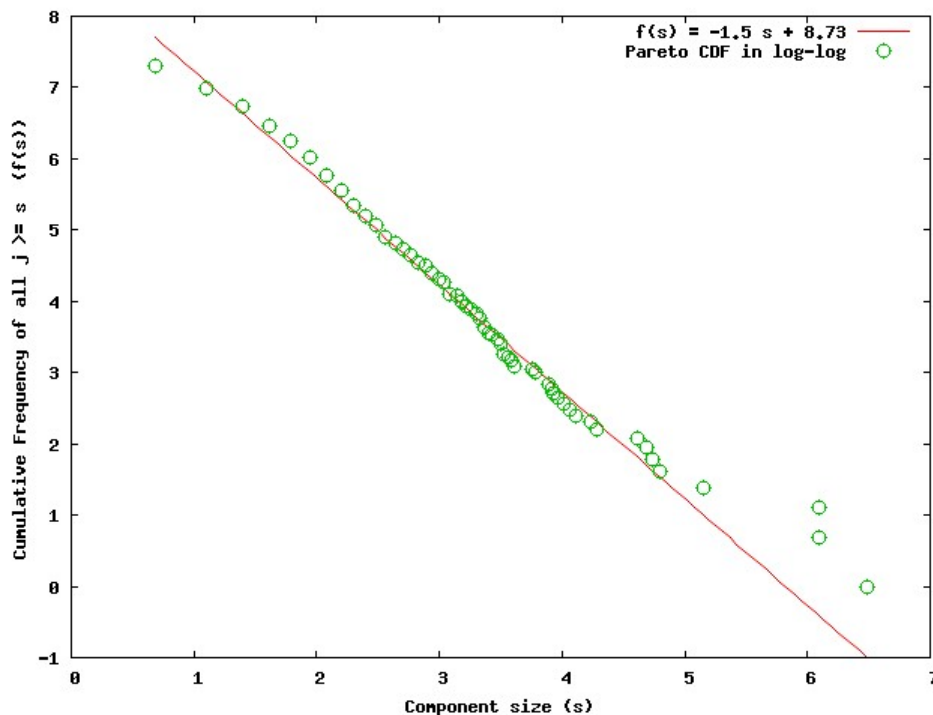


Figure 4.8: Pareto cumulative distribution (log-scale) of component sizes in the Web crawl ECG

Analysis of the CiteSeer ECG

In the CiteSeer dataset, 394,747 citations – i.e. 22.68% of all the citations in the dataset – have been endorsed by at least one co-citation. The number of articles upon which these endorsed citations are incident is 134,104 – i.e. 18.71% of the total number of articles in the dataset.

The co-citations that endorse citations in CiteSeer also exhibit a power-law in the number of co-citation counts, as shown in Figure 4.9, with an exponent of 3.53; the initial segment of this distribution deviates from the power-law.

Based on the distribution in Figure 4.9, we choose a co-citation count

threshold of 5 for the CiteSeer snapshot, since a very small number of citations have been endorsed by 5 or more co-citations. This indicates that these citations might genuinely connect topically relevant documents.

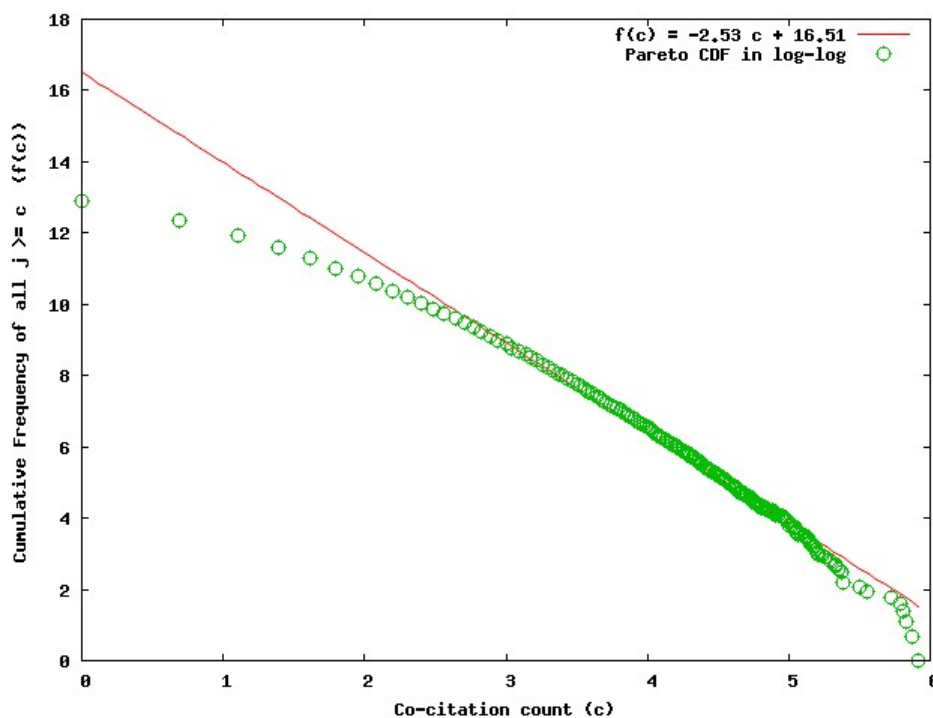


Figure 4.9: Pareto cumulative distribution (log-scale) of co-citation counts for endorsed citations in the CiteSeer snapshot

In CiteSeer, only 79,248 citations have been endorsed by 5 or more co-citations. This forms 20.08% of all the endorsed citations and only 4.55% of the total number of citations in the dataset. These citations, along with the articles on which they are incident, form our ECG. The number of articles in this ECG is 36,936.

As in the case of the Web crawl, in CiteSeer too, the endorsed citation data forms a very small percentage of the overall dataset. Hence, for the same

reasons discussed in the case of the Web crawl ECG, we believe that this is all the more reason to study semantics inherent in endorsed citations in the CiteSeer ECG. While the modal percentage of endorsed outgoing citations from the articles in the CiteSeer ECG is 50, with a probability of 0.745, the articles have 60% or fewer of their outgoing citations endorsed.

The indegree and outdegree distributions for this ECG are shown in Figures 4.10 and 4.11 respectively. Akin to the Web crawl ECG, the indegree distribution for the CiteSeer ECG too follows a power-law with a deviating initial segment, with an exponent of 3.3. However, unlike the Web crawl ECG, the outdegrees of the CiteSeer ECG seem to follow a power-law with an exponent of 5.49, albeit with an initial segment that deviates from the power-law towards a Poisson-like distribution.

Similar to the Web crawl ECG, in the CiteSeer ECG too, there are a large number of documents with only a few incoming endorsed citations, while there are only a few documents with a large number of incoming endorsed citations. The latter can be seen as topically authoritative documents to which several documents have endorsed citations.

In the CiteSeer ECG, the number of articles with a non-zero indegree is significantly smaller than the number of articles with a non-zero outdegree (58.49% and 81.68% respectively). However, while the maximum indegree here is 220, the maximum outdegree is only 40. Here too, the best topical hubs in the ECG do not point to a very large number of documents, but the top authorities are “recommended” very highly with large numbers of endorsed citations to them.

The CiteSeer ECG is a disconnected graph. Component sizes in the

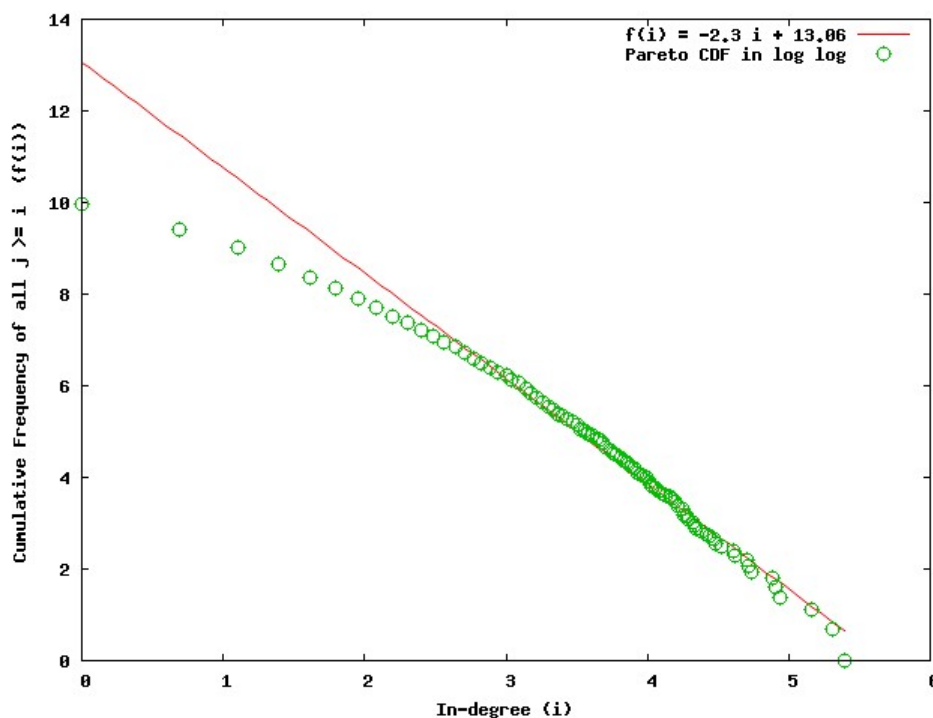


Figure 4.10: Pareto cumulative distribution (log-scale) of indegrees for the CiteSeer ECG

CiteSeer ECG, however, exhibit a significantly different behavior compared to the Web crawl ECG. Here, one of the components dominates over all the other components with 29,964 articles and 72,653 citations – i.e. 81.12% of the articles and 91.68% of the citations in the CiteSeer ECG. The distribution of the ECG component sizes is shown in Figure 4.12.

Since the ECG components are formed entirely out of endorsed citations, we look at them as being topical in nature – i.e. every document in a given component belongs to the same broad topic. This would explain the distribution of component sizes in the CiteSeer ECG. The ECG is made up largely of a single component because all the articles in it belong to the same topic.

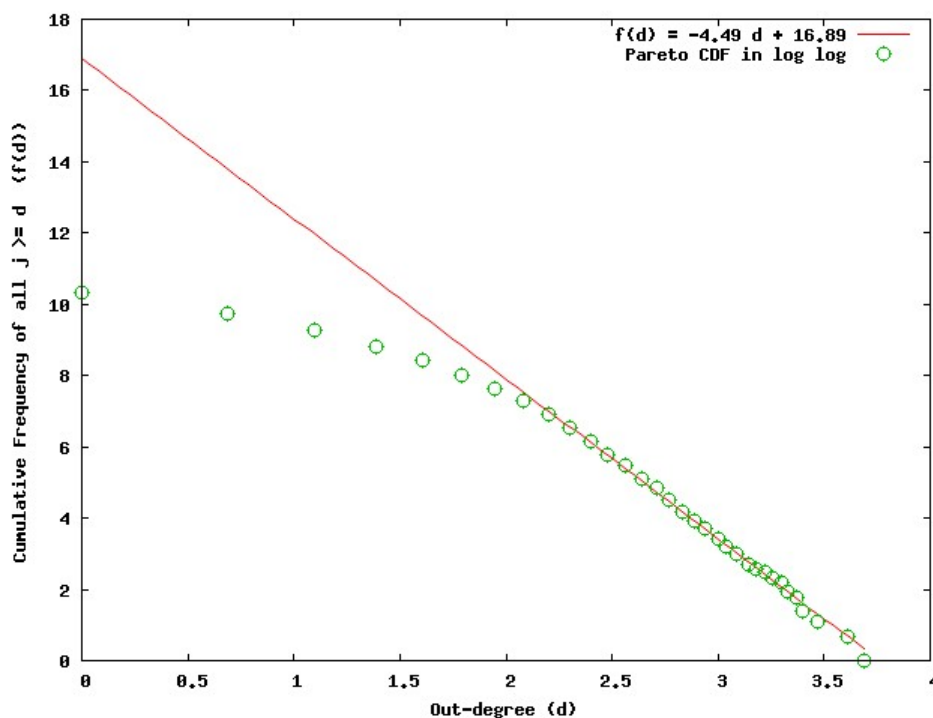


Figure 4.11: Pareto cumulative distribution (log-scale) of outdegrees for the CiteSeer ECG

This is by extension of the fact that the entire CiteSeer snapshot itself is on a single broad topic, viz. Computer and Information Sciences. This is in contrast to the Web crawl, which is made up of several heterogeneous topics.

It may be noted that a given topic could be distributed across two or more components in the ECG. The following are some of the most frequent terms across the titles of all the articles in the dominant component of the CiteSeer ECG: *system, network, data, analysis, distributed, algorithm, logic, model, learning, performance*. This list of terms is indicative of the topic of this component.

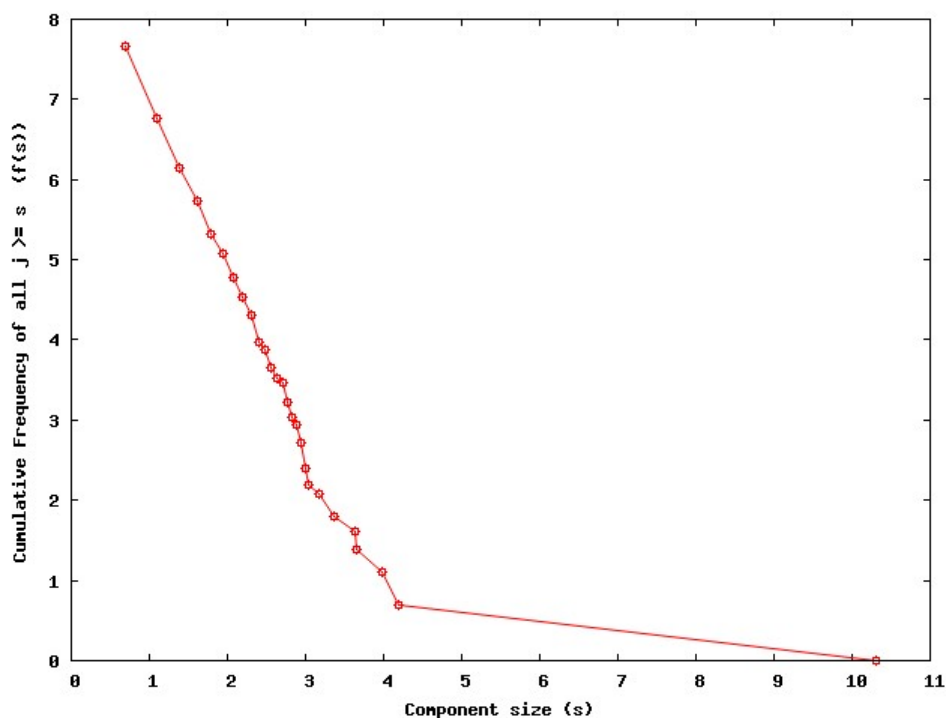


Figure 4.12: Pareto cumulative distribution (log-scale) of component sizes in the CiteSeer ECG

Structural Motifs

We now briefly examine some of the interesting structural motifs that are prevalent in both our ECGs. These are illustrated in Figure 4.13.

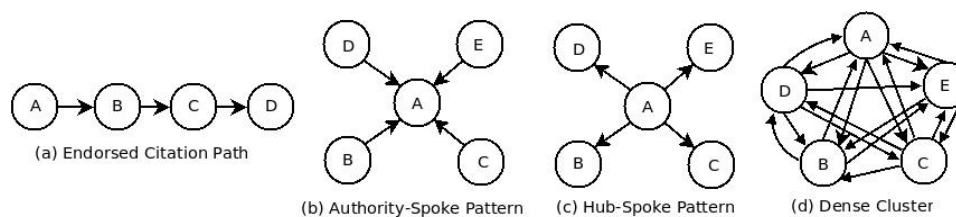


Figure 4.13: Some interesting structural motifs in the ECGs

Endorsed Citation Paths A commonly occurring structural motif in both

the ECGs is the endorsed citation path. This is a sequence of documents such that each intermediate document contains an outgoing endorsed citation to its successor document and an incoming endorsed citation from its predecessor document. The intermediate documents are deemed to have a common context with their predecessor as well as successor, thus allowing a topical transition from their predecessor to their successor.

Hubs/Authorities and Spokes We see several structures wherein a central “authoritative” document contains incoming endorsed citations from numerous spoke documents, which in turn do not often contain endorsed citations to one another. Here, the central document establishes a common context where the topics of the various spokes converge. In some cases, a central hub document contains *outgoing* endorsed citations to several other documents. The hub establishes a common point of departure for the topics of all the spokes. In a few cases, a single page acts as the hub as well as the authority around a set of spoke pages.

Dense Clusters We observed several subgraphs that had a dense structure. Many of these subgraphs were strongly connected, and sometimes cliques as well. Such a densely connected subgraph indicates that each document has found a common topical context with all or most of the other documents in that subgraph. Therefore, a single, strong topical context encompassing all or most of the documents is very likely to emerge in a dense ECG cluster.

One of the more curious patterns in our Web crawl ECG is two or more dense clusters being connected to one another through a “bridge page”. Figure 4.14 shows an actual ECG component with instances of this pattern.

A manual inspection of the pages of this component reveals that they contain news articles on various aspects of *Sports* in the online newspaper Cincinnati Enquirer¹⁰ over various months of the year 2004. Even though the component as a whole has the same theme (i.e. *Sports*), we see that there are several densely connected clusters making up this component. The way the clusters are chained in an almost linear fashion depicts the temporal nature of the topic.

We can see 11 different clusters in this component. We manually labeled these clusters as:

1. General Sports news for 30 March 2004
2. General sports news for 06 April 2004
3. News on the Cincinnati Reds baseball team for 06 April 2004
4. General Sports news for 26 April 2004
5. News on the Cincinnati Reds baseball team for 25 July 2004
6. General Sports news for 04 August 2004
7. General Sports news for 06 August 2004
8. General Sports news for 08 August 2004

¹⁰<http://www.enquirer.com>

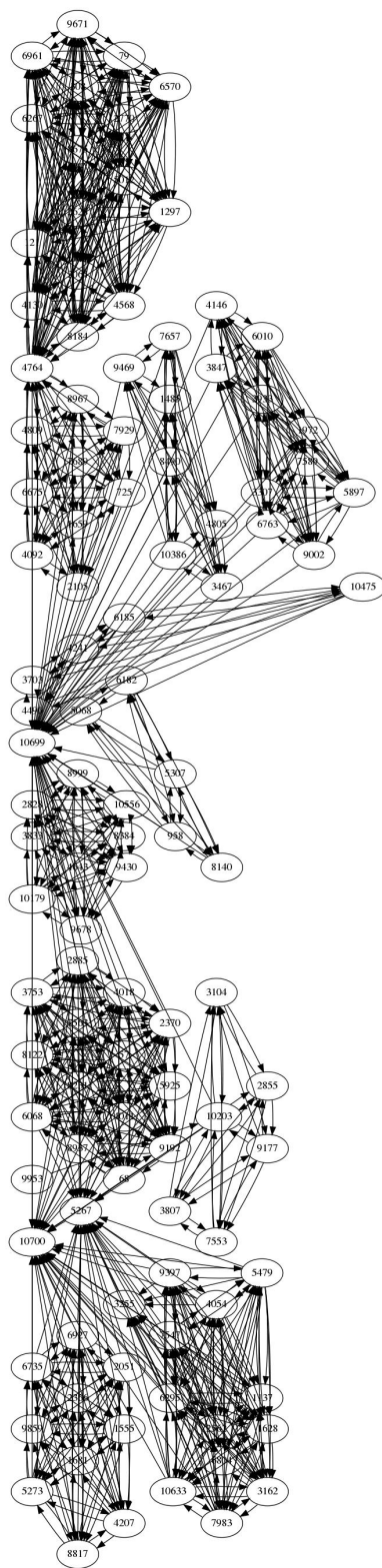


Figure 4.14: A component of the Web crawl ECG showing bridging of different dense clusters

9. General Sports news for 25 August 2004
10. General Sports news for 29 August 2004
11. General Sports news for 08 September 2004

The bridge pages establish a common ground for two or more clusters in such components. In our example above, the bridge connecting the various clusters for July and August 2004 is the page on the *2004 Summer Olympics Schedule*. Since the Summer Olympics were held from 13 August 2004 to 29 August 2004, there have been several news articles on it before and during the event, and these articles have linked to the games' schedule page. Thus, the common ground across all these news articles is established by the games' schedule page. This page is an example of a topically authoritative page aggregating the contexts of various clusters.

In this section, we have presented the notion of endorsed citations, and our experimental analyses of endorsed citations in a Web crawl and in CiteSeer. In Section 4.3, we describe a topical document ranking scheme, which uses endorsed citations.

4.3 Document Ranking using Endorsed Citations

As discussed in Section 4.2, the endorsement probabilities (ρ) of the outgoing citations from a given page can be viewed as a measure of the topical relevance of the corresponding out-neighbors to that page. Based on this idea, we propose a mechanism called *ERank* to rank documents within an ECG.

PageRank [Page et al., 1999] and Online Page Importance Computation (OPIC) [Abiteboul et al., 2003] consider a *random surfer* model, where the surfer, at each page, chooses any one of the outgoing citations on that page with a uniform probability. In our model, we consider a *topically biased surfer*, who, at each page, chooses an outgoing citation with a non-uniform probability – namely the endorsement probability (ρ) of the citation. However, the notion of PageRank cannot be directly applied to the ECG. The ECG is not only a set of disconnected components, but also any given component of the ECG is not guaranteed to be irreducible and aperiodic – necessary factors for the PageRank model.

In order to rank documents within an ECG component, we start by assuming that a topical surfer enters the component on any one of its pages with equal probability.¹¹ The surfer then crawls the component based on the endorsement probabilities of citations. The ERank of a page is then the probability that the topical surfer visits the page given that she has entered the component.

4.3.1 Formalization of ERank

We can formalize ERank as:

$$\mathbf{p}_{i+1} = L^T \mathbf{p}_i \quad (4.4)$$

Here, \mathbf{p}_i is the vector of ERank scores at iteration i , while L is a $M \times M$

¹¹Strictly speaking, the surfer need not start with pages within the ECG with equal probability. Pages in the ECG may be found non-uniformly through keyword-based search. However, we ignore this in order to understand topical surfing using ERank.

matrix (M is the number of pages within the component for which ERank is being computed) such that

$$L[u, v] = \begin{cases} \rho(u \Rightarrow v) & \text{if } (u, v) \in E' \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

It may be noted that the ERank vector is computed for each component separately, and not for the ECG as a whole. At the beginning of the ERank computation, each $\mathbf{p}_0[i]$ is initialized to $\frac{1}{m}$, where m is the number of nodes in the ECG component for which ERank is being computed.

We now take a brief detour and explain an iterative algorithm called *Online Page Importance Computation* (OPIC) [Abiteboul et al., 2003]. We use an OPIC-like model for implementing ERank. In the OPIC model, there are two vectors known as *cash* and *history*. The components of these vectors correspond to the pages on the Web. The cash of a node/page indicates the “votes” it has accumulated from its in-neighbors¹² in the previous iteration. To start with, each node is given the same amount of cash (typically $\frac{1}{n}$, where n is the number of nodes). Then, each node is picked infinitely often, whereupon it adds its current cash contents to its history, distributes all its cash among its out-neighbors uniformly, and resets its cash to 0. The history of a node, at the end of the computation, indicates the amount of cash that has flowed through it. This represents the “visit rate” or reachability of the node in a random walk. For practical purposes, the computation is stopped when the history vector reaches a steady-state distribution. The flow of cash

¹²The in-neighbors of a node A are the nodes that cite A .

in the OPIC model can be likened to the random surfer in PageRank. In fact, it is shown that the history vector of OPIC, like the PageRank vector, converges to the principal eigenvector of the Web graph. In order to make the Web graph irreducible and aperiodic, the OPIC model introduces a virtual node in the graph, to which all nodes have a link, and which in turn links to all nodes.

Due to its ease of implementation, we consider an OPIC-like cash-flow model for ERank. We assign an initial cash value to each of the nodes (i.e. pages) in an ECG component. The nodes are then randomly chosen infinitely often, and their cash contents are distributed among their out-neighbors based on the endorsement probabilities. This process is continued till the ranking of pages within the component based on the history of cash-flow becomes stable.

In addition to cash-flow based on endorsement probabilities, the other difference between ERank and OPIC is that the ERank model is allowed to “leak” cash out of the system. In other words, we do not use OPIC’s virtual node in ERank. Since the surfer is assumed to be a random surfer in OPIC, she is allowed to resume surfing from any page uniformly at random, once she encounters a page with no outgoing citations. In our case, however, we discard all the cash from nodes having zero outgoing citations, in effect allowing cash to leak out of the system. This is because a given component is not guaranteed to be strongly connected. This is also significant because the surfer in ERank is a *topically-biased surfer* unlike the random surfer in OPIC. The resultant distribution of ERank, therefore, need not correspond to the principal eigenvector of the component. However, it corresponds to some

stationary distribution, such that the ERank of a document corresponds to the topic-sensitive reachability of the document from anywhere in the component.

Topic-sensitive PageRank by [Haveliwala, 2003] also computes document importance values based on topicality. However, it relies heavily on topic taxonomies like the Open Directory Project (ODP).¹³ In our approach, we use endorsed citations to build topical backbones of the Web, thus eliminating the need to use topic taxonomies. Also, Topic-sensitive PageRank uses uniform probabilities for following outgoing citations from a given document. [Kleinberg, 1999] and [Lempel and Moran, 2001] also address topic-sensitivity. However, they too do not distinguish between outgoing citations.

We hypothesize that not all outgoing citations are equally relevant to the surfer, and that co-citations are a good measure for distinguishing their relevance. We have described above the idea of ERank as applied to the Web crawl ECG. However, this idea can be applied in the context of the ECG for any repository space, in general.

4.3.2 Experimental Analysis

A pertinent idea now is to contrast ERanks within an ECG component to the corresponding PageRanks. In order to accomplish this, we first needed to compute the PageRanks of the documents in our Web crawl as well as CiteSeer. Since computing PageRank using power iterations is highly resource-intensive, we estimated the PageRanks of documents as described by [Fortunato et al., 2007]. Across the entire citation graph, for each degree class

¹³<http://www.dmoz.org/>

$\mathbf{c} \cong (c_{in}, c_{out})$,¹⁴ where c_{in} and c_{out} are specific indegree and outdegree respectively, we computed the average PageRank of pages belonging to \mathbf{c} as:

$$PR(\mathbf{c}) = \frac{d}{N} + \frac{1-d}{N} \frac{c_{in}}{avg(N_{in})} \quad (4.6)$$

Here, d is the damping factor, set to 0.15 as with [Page et al., 1999]. N is the total number of documents in the dataset, while $avg(N_{in})$ is the average indegree over all the documents in the dataset. For the Web crawl, we have $N = 8,430,736$ and $avg(N_{in}) = 10.0182$. For the CiteSeer snapshot, we have $N = 716,772$ and $avg(N_{in}) = 2.428$.

It may be noted that PageRank can be computed using Equation 4.6 only in graphs with no degree correlation. [Nikoloski et al., 2005] confirm that scale-free graphs do not generate degree correlation. Hence, the above methodology for computing PageRank can be employed on both our datasets. [Fortunato et al., 2007] offer evidence that, even for graphs with weak degree correlations, the average PageRank of documents as shown in Equation 4.6 accurately mirrors the actual PageRank as computed traditionally (c.f. [Page et al., 1999]). They also show that the fluctuation between the average degree-class PageRank and the actual PageRank decreases as the indegree increases.

For our experiments, we needed the PageRanks of the documents in both our ECGs. The pages in the Web crawl ECG had a minimum non-zero indegree of 10 in the overall Web crawl, while the documents in the CiteSeer ECG had a minimum non-zero indegree of 5 in the overall CiteSeer snapshot.

¹⁴Here, the degree class \mathbf{c} , which is a tuple (c_{in}, c_{out}) , represents a class of documents having identical behavior in terms of indegrees and outdegrees in the citation graph.

We argue that the documents in both these ECGs had a “high” indegree in their respective overall citation graphs because: (i) both graphs are scale-free with only a small number of documents having a high indegree, and (ii) the number of documents having an indegree of at least 10 in the Web crawl, as well as the number of documents having an indegree of at least 5 in the CiteSeer snapshot, are both small (11.37% and 10.81% respectively). Therefore, the average PageRanks as computed through Equation 4.6 are reasonably accurate.

For each of our ECGs, for each component, we computed the ranking correlation between ERank and PageRank using the Kendall τ coefficient¹⁵ with adjustments for ties [Abdi, 2007]. In 790 of the 1,474 components of the Web crawl ECG, all values of either ERank or PageRank (or both) were equal, and hence the τ coefficient could not be computed for them accounting for values tied at the same ranks. This was true of just 70 components in the CiteSeer ECG. For the remaining components, we observed that the τ coefficients varied widely from -1 (i.e. complete disagreement between ERank and PageRank orderings) to $+1$ (i.e. complete agreement between ERank and PageRank orderings) in both the ECGs. We note that 40.94% of the components have a τ coefficient of at most $+0.4$ in the Web crawl ECG, while 28.79% of the components have a τ coefficient of at most $+0.4$ in the CiteSeer ECG. This indicates that a significant proportion of components show either a “low” or a negative correlation between ERank and PageRank in both the ECGs.

¹⁵Kendalls rank correlation. http://www.statsdirect.com/help/nonparametric_methods/kend.htm. Last accessed 23 May 2012.

In a separate experiment, for a few components of the Web crawl ECG, we also obtained an approximation of the PageRanks through the Google toolbar,¹⁶ and computed the τ correlation coefficients between these PageRanks and the corresponding ERanks. Here too, the τ coefficients varied significantly (from -0.154 to $+0.6486$).

The above experiments show that there is no consistent correlation between ERank and PageRank in either the Web crawl or CiteSeer. Our study suggests that ERank may sometimes offer a different ordering based on topicality, compared to PageRank. This implies that ERank may be able to add distinct value to the user's browsing experience in terms of relevance. Hence, it is worthwhile to explore further the notion of endorsed citations for focused resource discovery and fine-tuned relevance ranking in repository spaces.

This concludes the second part of this thesis, where we have discussed semantics mining from citations in repositories. In the third part, we look at detecting semantic attributes of concepts in social spaces, as described in the next chapter.

¹⁶<http://toolbar.google.com>

5

Attribute Detection in Social Spaces

A *concept* in a social space is an abstract representation of an idea or an experience. A concept could mean a topic, theme or simply a named entity. A concept may encapsulate a set \mathbf{A} of other concepts as “attributes”. A set of concepts \mathbf{A} is called the set of *attributes* of a concept C (denoted by $attrs(C)$) if: (i) every concept $A \in \mathbf{A}$ describes one or more properties of

C , and (ii) the concepts in \mathbf{A} collectively describe C uniquely.¹ In this part of the thesis, we refer to a concept that is uniquely described by a set of attributes as the *object* of those attributes. The set of attributes of an object collectively lends a *commonsense meaning* to that object.

Given an attribute $A \in \text{attrs}(C)$, the property of C that A describes could be in the form of a semantic association between A and C . Also, since a set of concepts \mathbf{A} is defined as an attribute-set based on its collective ability to describe another concept, it is difficult to tell whether a given concept A is an attribute of another concept C if it is presented in isolation. Such a concept A could have a well-defined semantic association with C , but its ability to collectively, along with other concepts, uniquely describe C would not be clear.

For instance, given the object term *India*, we seek to mine concepts such as *New Delhi*, *Subcontinent*, *Mahatma Gandhi*, *Himalayas*, *Ganga*, etc. In this example, each of the attributes has a semantic association with *India*: (i) *New Delhi is the capital of (or is contained in) India*, (ii) *Subcontinent is a super-class of India*, (iii) *Mahatma Gandhi is the father of the nation of India*, (iv) *Himalayas is a geographical feature of India*, (v) *Ganga is a river of India*. Collectively, they help in uniquely describing the concept *India*, and therefore constitute its attribute set. However, if we consider the concept *Mahatma Gandhi* in isolation, we cannot be sure if it is describing *India*, as it is also associated with several other concepts such as *Satyagraha*, *South Africa*, etc.

¹Since attributes uniquely define concepts, for any two concepts C_1 and C_2 , $C_1 \neq C_2 \Rightarrow \text{attrs}(C_1) \neq \text{attrs}(C_2)$.

Also, not all topically relevant terms of an object are attributes of that object. For instance, the synonyms of a term provide *alternatives* to using that term (e.g., *decoration*, *adornment*, *embellishment*, etc.). However, they do not *describe* the concept represented by that term. Similarly, the semantic siblings of an object (c.f. [Brunzel, 2008; Brunzel and Spiliopoulou, 2007; Rachakonda et al., 2012]) do not help in uniquely describing that object. They represent other concepts that are semantically “equivalent” to the object (e.g., *Rahul Dravid*, *Sachin Tendulkar*, *Sourav Ganguly*, etc.). Now, since not all topically relevant terms of an object are the attributes of that object, we assert that the task of identifying object-attribute relationships is not the same as the task of topic modeling for the object (c.f. [Anthes, 2010; Blei and Lafferty, 2007; Blei et al., 2003; Hofmann, 1999a]).

We address the problem of detecting the attributes of a given object as held by the global world-view of a population. We propose two co-occurrence based hypotheses for detecting such object-attribute relationships in social spaces. We have tested our hypotheses with Wikipedia as the underlying social space.

Dataset For our experiments, we used the December 2006 dump of the English Wikipedia. We treated each section within a Wikipedia article as an occurrence context (or a socio-cognitive context) within which concepts co-occur. We considered only non-stub articles that contain at least one section with at least two terms in it. We treat terms within the article-sections as concepts in this social space. Given an article-section, we consider as “terms” those phrases that form the titles of the target articles of hyperlinks within

that article-section. In all, we considered 3, 217, 187 Wikipedia articles. The dataset contained 9, 145, 712 article-sections (i.e. socio-cognitive occurrence contexts) and 5, 687, 833 terms (i.e. concepts).

We conducted all the experiments described in this chapter on a computer having a 4-core Intel Xeon processor with IA-64 architecture, 6 GB of RAM and a 200 GB hard disk drive. We implemented our algorithms in Perl, with MySQL as the database for storing our dataset.

We now discuss the modeling of co-occurrence patterns of concepts in a social space, which will enable us to formulate our hypotheses for attribute detection.

5.1 Co-occurrence based Attribute Detection

Let D be the set of documents (resulting from SCPs) in the social space, while T is the set of all terms in the social space. An occurrence context c is characterized by a set of terms $c \in 2^T$. The terms in an occurrence context are said to have *co-occurred* within that context. Patterns of such co-occurrences form the basis of object-attribute semantics and other semantic associations. The set of all occurrence contexts of a document $d \in D$ is denoted by C_d . The set of all occurrence contexts across the entire social space is denoted by C_D . The set of all occurrence contexts containing a given term t is denoted as C^t . Based on the above notions, we now describe a set of primitives by which we can make inferences about observed patterns of object-attribute relationships across documents.

Given a term t , its *support* is defined as the ratio of the number of contexts

it appears in, to the total number of contexts. This is given by

$$\text{sup}(t) = \frac{|C^t|}{|C_D|} \quad (5.1)$$

Terms are not uniformly used along with one another. Some terms tend to be used together much more than some other terms within an occurrence context. *Usability* is a measure of the probability of a term being used in the context of some other term. The usability of a term v in the context of a term u is given by

$$\rho(v|u) = \frac{|C^u \cap C^v|}{|C^u|} \quad (5.2)$$

When a set of terms is to be considered, we also use two other primitives using which two canonical forms of their combined semantics can be expressed: *closure* and *focus*. Given a set of terms σ , their closure is the set of all occurrence contexts containing at least one of the terms in σ , and their focus is the set of all occurrence contexts containing all the terms in σ . Closure is denoted by σ^* , and focus is denoted by σ_\perp .

$$\sigma^* = \bigcup_{u \in \sigma} C^u \quad (5.3)$$

$$\sigma_\perp = \bigcap_{u \in \sigma} C^u \quad (5.4)$$

5.1.1 Modeling Co-occurrence Patterns of Concepts

The social space corpus is visualized as an *occurrence graph*, which is a bipartite graph mapping terms to their occurrence contexts. From the occurrence

graph, different higher forms of graphs are generated. These are the *co-occurrence graph* and the *usability graph*. The bipartite occurrence graph, denoted by \mathcal{O} , captures the association between terms and their occurrence contexts. Formally the occurrence graph is defined as:

$$\mathcal{O} = (\{T, C\}, E) \quad (5.5)$$

where the set of nodes is partitioned into T , the set of terms, and C , the set of occurrence contexts. $E \subseteq T \times C$ is the set of associations between the terms and the occurrence contexts.

Two terms t_1 and t_2 are said to have *co-occurred* if there exists at least one occurrence context c that contains both t_1 and t_2 . Thus, the entire corpus can be represented as a *co-occurrence graph*. Edges between terms in this graph represent the pair-wise co-occurrences across terms. The co-occurrence graph is the basic data structure, on top of which we design hypotheses for object-attribute relationships. Formally, the co-occurrence graph is defined as:

$$G_{\mathcal{O}} = (T, E, w) \quad (5.6)$$

where T is the set of all terms in the corpus, E is the set of all pair-wise co-occurrences across terms, and the function $w : E \rightarrow \mathbb{N}$ indicates the co-occurrence count between pairs of terms. The function w is given by $w(t_1, t_2) = |C^{t_1} \cap C^{t_2}|$.

Given a term t , its co-occurrence neighborhood, denoted by $N(t)$, is the set of all terms co-occurring with t . Formally, given a co-occurrence graph

G as defined in Equation 5.6 above, and a term $t \in T(G)$, we have

$$N(t) = \{v \mid \{t, v\} \in E(G)\} \tag{5.7}$$

The neighborhood $N(t)$ of a single vertex t can basically be visualized as the set of vertices of the “star”-like subgraph originating from t .

Given a co-occurrence graph $G = (V, E, w)$, the *semantic context* of a term t , denoted by $\mathbb{S}(t)$ or by $t(G)$, is the induced sub-graph of the vertices of its neighborhood. An induced subgraph H of a graph G contains a subset of vertices of G and all edges of the form $\{v_1, v_2\}$ from G such that $v_1, v_2 \in V(H)$. In other words

$$\mathbb{S}(t) = t(G) = (\{t\} \cup N(t), E_t, w_t) \tag{5.8}$$

where $\{v_1, v_2\} \in E_t \Rightarrow ((v_1, v_2 \in \{t\} \cup N(t)) \wedge (\{v_1, v_2\} \in E))$. Here, $w_t : E_t \rightarrow \mathbb{N}$ indicates the co-occurrence counts between terms (as defined earlier).

A semantic context can also be obtained for a *set* of terms. Given a co-occurrence graph $G = (V, E, w)$, the semantic context of a set of terms Q is defined as the induced subgraph of the vertices in $\bigcup_{u \in Q} N(u)$. This is denoted by $\mathbb{S}(Q)$.

The semantic context is an important primitive for mining latent semantics. We claim that the attributes of an object t will be found within the semantic context of t . It is not necessary to process the entire co-occurrence graph for extracting object-attribute relationships pertinent to a given term.

We conjecture that the same applies to the extraction of a large number of other latent semantics.

We define a set of terms X to be *coherent* if

$$\bigcap_{x \in X} N(x) \neq \phi \quad (5.9)$$

For the purpose of mining object-attribute relationships, a higher form of graphs, known as the *usability graph* is generated by using the occurrence graph and the co-occurrence graph. The usability graph of a given semantic context \mathbb{S} over a co-occurrence graph \mathcal{G} is a directed graph, whose nodes are terms from \mathbb{S} , and directed edges represent the usability scores between ordered pairs of terms within \mathbb{S} . The usability graph for \mathbb{S} is denoted by the adjacency matrix $U_{\mathbb{S}}$. This is discussed in detail further on in this chapter.

5.1.2 The Notion of Attribute Detection

If a set of concepts $Q \subseteq N(x)$ are observed, they are said to be the attributes of x if they collectively maximize the probability of guessing the object of discussion to be x . Intuitively, this hypothesis can be explained using the analogy of the popular parlor game called *Twenty Questions*. In this game, the objective is to determine an object (i.e. a person, a place or a thing) by asking at most twenty questions (about the attributes of the object), which we think maximize our chances of guessing the object correctly. The idea here is to optimally select the twenty attributes such that the probability of determining (or the “determinability” of) the object is maximized.

The above hypothesis about attributes can be expressed in terms of the

co-occurrence graph as follows: Given a term x , its attributes are elements of the coherent set of terms Q , such that x has the highest usability score with respect to Q . This can be formalized as:

$$Q = \arg_{A \in 2^{N(x)}} \max \rho(x|A) \quad (5.10)$$

However, we have proved that finding the set Q of attributes of x , which has the highest “determinability” for x (i.e. the highest usability of x with respect to Q), is an NP-hard problem. Please refer Appendix A for details of this proof. Since the determinability hypothesis is NP-Hard, we propose two other hypotheses for attribute detection in social spaces. The first such hypothesis is called the Usability Hypothesis, which is described in Section 5.2.

5.2 The Usability Hypothesis

Given a coherent set of terms Q in a corpus represented as a co-occurrence graph G , a term $u \in V(\mathbb{S}(Q))$ is an attribute of a term $v \in V(\mathbb{S}(Q))$ if $u \in N(v)$ and v has a higher usability score than u in an infinitely long random walk executed on $\mathbb{S}(Q)$. The notion of *usability* – i.e. $\rho(y|x)$ – represents what might be termed the probability of “using” y in a context mentioning x . The pairwise usability scores $\rho_Q(y|x)$ between any pair of nodes (x, y) in $\mathbb{S}(Q)$ are computed using the closure of Q as the underlying corpus:

$$\rho_Q(y|x) = \frac{|Q^* \cap C^x \cap C^y|}{|Q^* \cap C^x|} \quad (5.11)$$

In other words, $\rho_Q(y|x)$ is the probability that an actor (or content creator) uses y “through” x in an overall context pertaining to Q .

For example, $\rho_{Barack\ Obama}(Tenure\ of\ Office|US\ President)$ is the probability of using the term *Tenure of office* in a context mentioning *US President*, among the set of all contexts pertaining to *Barack Obama*. In this sense, $\mathbb{S}(Q)$ can be visualized as a *usability graph* – denoted by its adjacency matrix $U_{\mathbb{S}(Q)}$ – in the overall context of Q , where the weight of a directed edge from a node u to another node v is given by $\rho_Q(v|u)$.

The Usability Hypothesis may be stated as follows: *In the context of a set of terms Q , the usability of an object is more than the usability of its attributes. Further, it is more likely that the attribute is used in an occurrence context that mentions the object, rather than vice-versa.*

For instance, consider the set of all contexts relating to “Barack Obama”. Consider “Tenure of Office” to be an attribute of “US President”. According to the Usability Hypothesis, if an actor were to be randomly creating contexts mentioning “Barack Obama”, she would be more likely to use the term “Tenure of Office” after using the term “US President”. This is because the term “US President” – and not “Tenure of Office” – has a higher usability in the context of the term “Barack Obama”. Further, it is more likely that the actor uses “Tenure of Office” in an arbitrary context mentioning “US President” rather than using “US President” in an arbitrary context mentioning “Tenure of Office”. Since there are different ways by which one can use either of these terms, the overall usability of a term in the neighborhood is calculated by a fixed point computation over a random walk. We call this process *UseRank*.

5.2.1 Usability Ranking

If $V(\mathbb{S}(Q))$ is represented by the vector \mathbf{p}_Q and if $U_{\mathbb{S}(Q)}$ is the corresponding adjacency matrix of the usability graph defined over $\mathbb{S}(Q)$, where $U_{\mathbb{S}(Q)}[u, v] = \rho_Q(v|u)$, the UseRank of nodes can be formalized as follows:

$$\mathbf{p}_Q = U_{\mathbb{S}(Q)}^T \mathbf{p}_Q \quad (5.12)$$

The UseRank vector \mathbf{p}_Q converges to some fixpoint of $U_{\mathbb{S}(Q)}^T$.

In terms of its functioning, UseRank can be likened to cash-flow based ranking models such as Online Page Importance Computation (OPIC) [Abiteboul et al., 2003]. Initially, each node in $\mathbb{S}(Q)$ is given an equal amount of cash. Each node also maintains a history of the cash that has flowed through it over various iterations. Each node is picked infinitely often to perform the following action: the node updates its history with its current cash contents (i.e. $history(u)+ = cash(u)$), and distributes its current cash among its outneighbors proportionate to their usability probabilities w.r.t. itself. Note that instead of being conserved, cash is synthesized during this process. This is because the UseRank of a node is a measure of its “usability” rather than its “visitability”. At the end of this iterative history updating and cash distribution process, the history values of the nodes represent the usability of their corresponding topics.²

A measure of “topical impact” is also introduced into UseRank computations where each node in $\mathbb{S}(Q)$ updates its history proportionate to its

²Another difference between OPIC and UseRank is that the underlying graph in OPIC may have some nodes with zero outlinks, whereas the usability graph defined over $\mathbb{S}(Q)$ has no such nodes. The usability graph defined over $\mathbb{S}(Q)$ is a strongly connected graph, meaning that every node has at least one child-node to transfer its cash to.

impact in the context of the topic of Q . In other words, $history(u)_+ = TopicalImpact_Q(u) \times cash(u)$.

We define the topical impact of a node as:

$$TopicalImpact_Q(u) = \frac{|Q^* \cap C^u|}{|Q^*|} \quad (5.13)$$

The topical impact of a term u in the context of a set of terms Q can also be seen as the *usability* of u in the context of Q . The higher the topical impact of u , the more it is deemed to be relevant in the topic of Q . The notion of topical impacts thus allows for the updating of the history of a node proportional to its relevance in the context of the query terms.

Incorporating topical impacts into the UseRank model, Equation 5.12 becomes

$$\mathbf{p}_Q = U_{\mathbb{S}(Q)}^T E_Q \mathbf{p}_Q \quad (5.14)$$

where E_Q is a diagonal matrix such that $E_Q[u, u] = TopicalImpact_Q(u)$

In this variant, the UseRank vector \mathbf{p}_Q converges to some fixpoint of $U_{\mathbb{S}(Q)}^T E_Q$.

5.2.2 Object-Attribute Trees

The UseRank scores, as computed above, are used for identifying object-attribute relationships. Using the object-attribute relationships discovered in $\mathbb{S}(Q)$, we envisage a hierarchical structure of object-attribute relationships known as a *Object-Attribute Tree* (OAT). In an OAT on the topic represented by Q , an attribute concept can further have attributes and so on. For exam-

ple, in the context of the topic of “Barack Obama”: (i) *Barack Obama* has an attribute called *US President*, (ii) *US President* has an attribute called *Tenure of Office*, and so on. Figure 5.1 shows an example OAT.

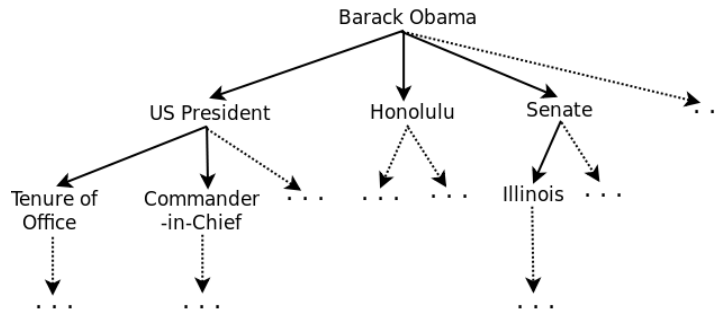


Figure 5.1: A hypothetical OAT in the semantic context of *Barack Obama*

In order to facilitate the building of such an OAT for $\mathbb{S}(Q)$, we hypothesize that a node u is an attribute (or child) of a node v if the following conditions are true.

- $u \in N(v)$. The idea here is that the attributes of an object co-occur with that object.
- $UseRank(v) \geq UseRank(u)$ following the usability hypothesis
- *Among all the neighbors of u in $\mathbb{S}(Q)$, v has the highest $UseRank$.* We impose this condition for simplicity – in order to avoid having multiple parents for a given concept in the OAT.
- *u is not an ancestor of v already.*³ This is because an OAT should not have cycles.

³It is plausible that v has already been assigned as an attribute to u – if u, v , and all the neighbors of v have the same $UseRank$.

Assigning parents to nodes in an OAT according to the above conditions, we find at least one node to which a parent cannot be assigned. If there is only one such node without a parent, we designate it as the root concept node and return the OAT rooted at this node. However, if there are multiple such root candidates, two strategies can be followed.

1. Designate each one of those root candidates as root concepts, and return the respective trees rooted at those nodes. In this case, we will have a forest of OATs rather than a single OAT.
2. Break the tie by picking the root candidate with the highest UseRank. If multiple root candidates qualify, pick one uniformly at random. We used this strategy in our experimental evaluations.

5.2.3 Experimental Analysis

We implemented UseRank with TopicalImpact on the Wikipedia dataset described earlier in Chapter 5. Given a term x input by the user, the set Q^* , where $Q = \{x\}$, was constructed as the set of article-sections containing at least 7 occurrences of x . The reason for this is as follows. On the average, an article-section contains only 2.28 named entities occurring 7 or more times. We speculate that these “small” number of entities may be the central concepts of these article-sections. Hence, we assume that Q^* is the set of article-sections where the query entity is a central concept. This is done in order to avoid topical contamination of $\mathbb{S}(Q)$ due to “topic drift”.

For practical purposes, $\mathbb{S}(Q)$ was constructed using only the 10 most frequently occurring entities for each article-section in Q^* . We then ran our

algorithm over $\mathbb{S}(Q)$ for 25 iterations.

For a given query, we computed the OAT and listed the top-10 attributes of the root concept (i.e. the top-10 children of the root concept ordered by UseRank scores). If more than one concept qualified for being the root concept, we chose the one with the highest UseRank score. If more than one root concept had the highest UseRank score, we broke the tie by picking a root concept randomly.

We issued 30 queries chosen by 10 volunteers, and for each of these queries, we presented the root concept and its top-10 attributes to the volunteers. We asked the volunteers to evaluate each query on the following criteria.

1. *Is the root concept relevant to the query?* This is an essential criterion for evaluation, since the attributes of any entity within the OAT are said to hold within the overall context of the root concept. The root concept is computed in an emergent fashion in this approach.
2. *Is the root concept appropriate to the query, or is it too general or too specific?* Sometimes, the root concept computed by our algorithm may establish an overall context that may be too general or too specific in response to a query. For example, in response to the query “Himalayas”, picking “Tibet” as the root concept may be seen as being too specific, since “Tibet” is not sufficient to establish an overall context within which to describe the “Himalayas”. We asked the volunteers to assign points to the root concept ranging between -5 and $+5$, where a negative score indicates generality and a positive score indicates specificity (e.g., -5 stands for “extremely general” and $+5$ stands for “extremely

specific”). A score of 0 indicates that the root concept is the same as the query.

3. *Of the 10 attributes shown for this root concept, how many do you agree with?* In other words, we asked the volunteers to identify a subset of these 10 attributes, which they think help in uniquely describing the root concept. Using this information, we measured the precision of the computed attribute set for a given root concept, for a given volunteer.

The trial queries used in our experiments are shown in Table 5.1.

The following are two samples of the top-10 attribute sets generated by the Usability Ranking approach for the root concepts of given queries.

Root concept *infosys*: The top-10 attributes are *murti, pune, bangalore, india, n. s. raghavan, mysore, share, technology, city-tv* and *infosys bpo limited*.

Root concept *google*: The top-10 attributes are *map, page, yahoo!, states, gmail, app, android, web search engine, china* and *news*.

Relevance of the Root Concept to the Query

For 29 of the 30 trial queries, all 10 volunteers agreed that the root concept was relevant to the query. For the remaining query, 9 volunteers agreed that the root concept was relevant to the query, while the tenth volunteer disagreed. This shows that the root concept computed by the UseRank algorithm is relevant to the given query in most cases.

Sl. No.	Query	Sl. No.	Query
1	himalayas (tibet)	2	infosys
3	bangalore	4	mohandas karamchand gandhi
5	mayawati	6	red fort
7	taj mahal (mahal)	8	jawaharlal nehru
9	lalu prasad yadav (yadav)	10	bay of bengal
11	mahendra singh dhoni	12	nepal
13	facebook	14	google
15	latin america (america)	16	michael jackson (jackson)
17	mysore	18	karnataka
19	mount everest	20	biryani
21	turing machine	22	digital camera
23	japanese cuisine	24	siberian tiger
25	african elephant	26	demographics (trip distribution)
27	christopher nolan	28	motherboard
29	manmohan singh	30	popeye (fleischer)

Table 5.1: List of queries used for evaluation of the usability ranking approach. Queries 1, 7, 9, 15, 16, 26 and 30 had root concepts different from the query. The root concepts of these queries have been mentioned in parentheses.

Generality/Specificity of the Root Concept

For each query, say q , we computed the average generality/specificity, a_q , over all 10 volunteers. Figure 5.2 shows the frequency distribution of the a_q scores (rounded off to the nearest integer) for our trial queries. The a_q scores range from -3.7 to $+3.8$, with 23 of the 30 root concepts having an a_q score of 0.

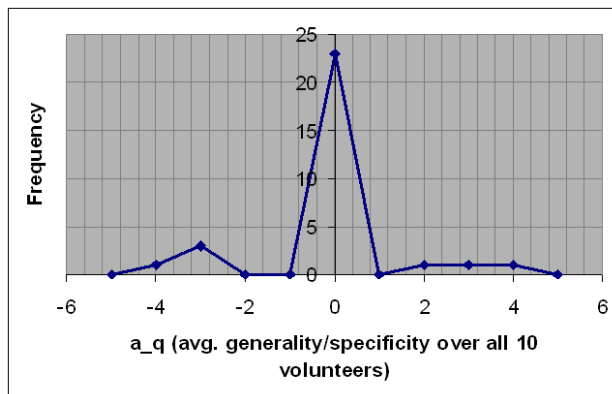


Figure 5.2: UseRank: Frequency distribution of the a_q scores (rounded off to the nearest integer) for the 30 trial queries

We then computed a consolidated score of generality/specificity, g , as the average of the a_q scores over all 30 queries. We found that $g = -0.05$, thus indicating that the root concept computed by the UseRank algorithm is neither too general nor too specific in most cases.

Precision of Attributes for the Root Concept

For each root concept, say q , we computed the precision of attributes for each volunteer, and then averaged the precision over all 10 volunteers. These average scores, p_q , ranged from a minimum of 0.42 to a maximum of 0.92. Figure 5.3 shows a plot of the p_q scores for each of the trial queries. We observe that 13 of the 30 queries had a p_q score of 0.7 or more. Figure 5.4 shows the frequency distribution of the p_q scores (rounded off to one decimal place) for our trial queries.

Over all 30 queries, the average of the p_q scores was found to be 0.66 with a standard deviation of 0.12. These results show that the usability

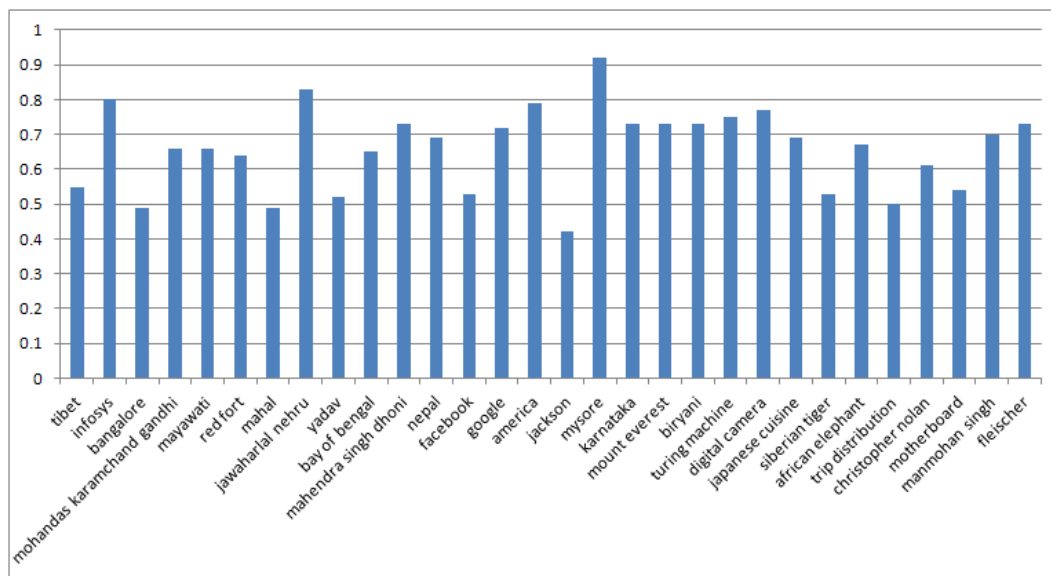


Figure 5.3: UseRank: p_q scores for the root concepts for the 30 trial queries. The X-axis represents the trial queries, and the Y-axis represents the p_q scores.

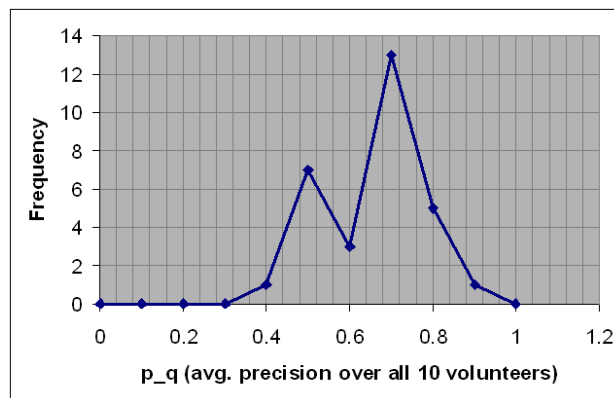


Figure 5.4: UseRank: Frequency distribution of the p_q scores (rounded off to one decimal place) for the 30 trial queries

ranking approach has the potential to address the problem of identifying object-attribute relationships for building OATs.

In Section 5.3, we describe another hypothesis, which could be used

in augmentation with the usability ranking approach, for detecting object-attribute relationships in social spaces.

5.3 The Positive Relevance Hypothesis

The next hypothesis that we propose for detecting object-attribute relationships is known as the Positive Conditional Relevance Hypothesis, based on the probabilistic dependence between two terms. Given a co-occurrence graph G , we define a term y to be an attribute of a term x if the usability of y with respect to x is greater than the support of y .

Given two terms x and y , we say that their occurrences (or usages) are independent of one another if $\rho(y|x) = \text{sup}(y)$ (or alternatively, if $\rho(x|y) = \text{sup}(x)$). In other words, if the probability of using y in a context containing x , is the same as the probability of using y in any arbitrary context, then the usages of x and y within the text corpus are said to be independent of one another. However, x and y are probabilistically dependent if $\rho(y|x) \neq \text{sup}(y)$.

If $\rho(y|x) > \text{sup}(y)$, it is said that there is a *positive relevance* of y to x , while $\rho(y|x) < \text{sup}(y)$ indicates *negative relevance* of y to x [Falk and Bar-Hillel, 1983]. The positive relevance of a term y to a term x indicates that the usability of y is highly conditional to the usage of x . In this approach, we simply consider the usability of y with respect to x – i.e. $\rho(y|x)$ – for every $y \in N(x)$. Given this, the Positive Relevance Hypothesis may be stated as follows: *An actor (content creator) is more likely to use an attribute of x while creating some contexts mentioning x , rather than using it independently by itself.*

We hypothesize that the usage of the attribute of an object is highly positively relevant to the usage of the object itself – i.e. $\rho(y|x) > \text{sup}(y)$ if y is an attribute of x . For instance, given that the term “Tenure of Office” is an attribute of the term “US President”, according to the Positive Relevance Hypothesis, an actor creating contexts mentioning “US President” is more likely to use “Tenure of Office”, than an actor randomly creating a context.

5.3.1 Quantifying Positive Relevance

Given that x is the object for which attributes are to be detected, we compute a *Relevance Score*, r_y , for each $y \in N(x)$, as $r_y = \rho(y|x) - \text{sup}(y)$. We hypothesize that a term y is an attribute of x if $r_y > m$, such that $m \in [0, 1)$. Here, m is a tunable parameter. The higher the value of r_y , the higher the “attributeness” of y to x .

For practical purposes, we do not hard-code the parameter m . Instead, we list the terms in $N(x)$ in the descending order of their r_y scores, and present the top- k such terms (along with their r_y scores) as candidate attributes of x to the social space analyst.

5.3.2 Experimental Analysis

We tested this approach too using the Wikipedia dataset described earlier. For each of the 30 queries chosen by volunteers for our previous approach, we obtained the top-10 attributes using the Positive Relevance approach.

The following are two samples of the top-10 attribute sets generated by the Positive Relevance approach for given queries.

Query bay of bengal: The top-10 attributes are *india, river, sea, bangladesh, bengal, big five personality traits, indian ocean, orissa, andhra pradesh* and *burma*.

Query mysore: The top-10 attributes are *india, karnataka, bangalore, kannada language, tippu sultan, wodeyar, empire, kingdom of mysore, acer laboratories incorporated* and *mangalore*.

A team of 10 volunteers was asked to evaluate each of our queries for the precision of its attributes. In other words, we asked the volunteers to identify a subset of these 10 attributes, which they think help in uniquely describing the query concept. For each query, say q , we computed the precision of attributes for each volunteer, and then averaged the precision over all 10 volunteers. These average scores, p_q , ranged from a minimum of 0.39 to a maximum of 0.87. Figure 5.5 shows a plot of the p_q scores for each of the trial queries, while Figure 5.6 shows the frequency distribution of the p_q scores (rounded off to one decimal place).

Over all 30 queries, the average of the p_q scores was found to be 0.65 with a standard deviation of 0.13. These results show that, along with the Usability Ranking approach, the Positive Relevance approach too could be useful in building OATs.

5.4 Usability Ranking vs Positive Relevance

Figure 5.7 gives a comparative view of the p_q scores for both our approaches, for some of the trial queries. In the Usability Ranking approach, even though

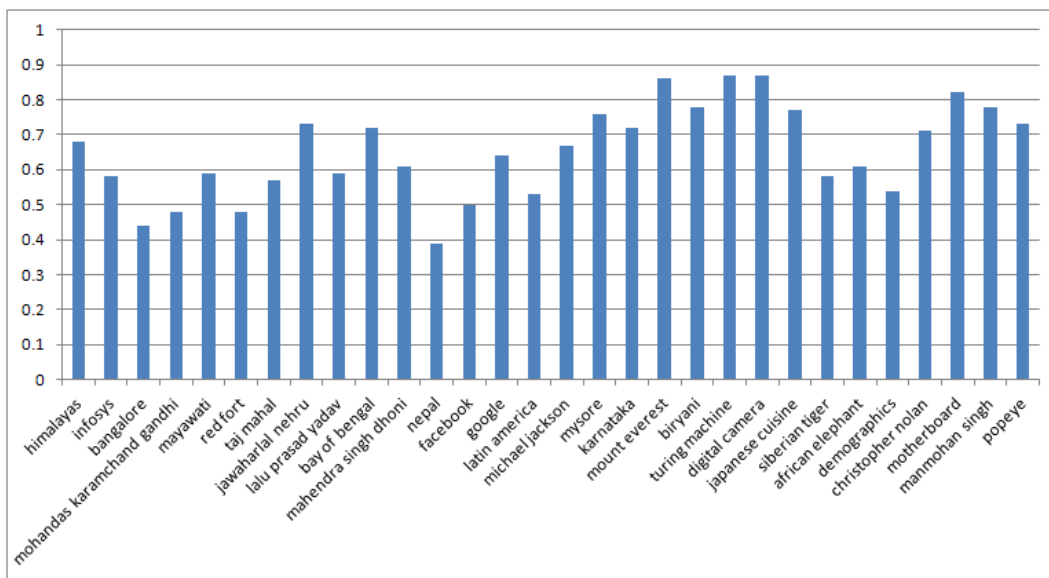


Figure 5.5: Positive Relevance: p_q scores for the 30 trial queries. The X-axis represents the trial queries, and the Y-axis represents the p_q scores.

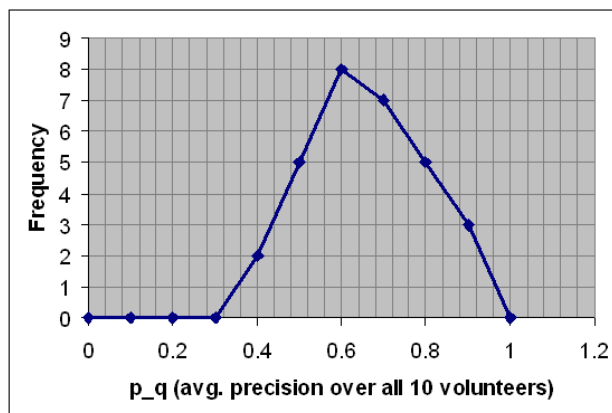


Figure 5.6: Positive Relevance: Frequency distribution of the p_q scores (rounded off to one decimal place) for the 30 trial queries

the attributes have been computed for the *root concepts* of the queries rather than the queries themselves, the root concept is the same as the query for 23 of the 30 queries. We have chosen these 23 queries for this comparison.

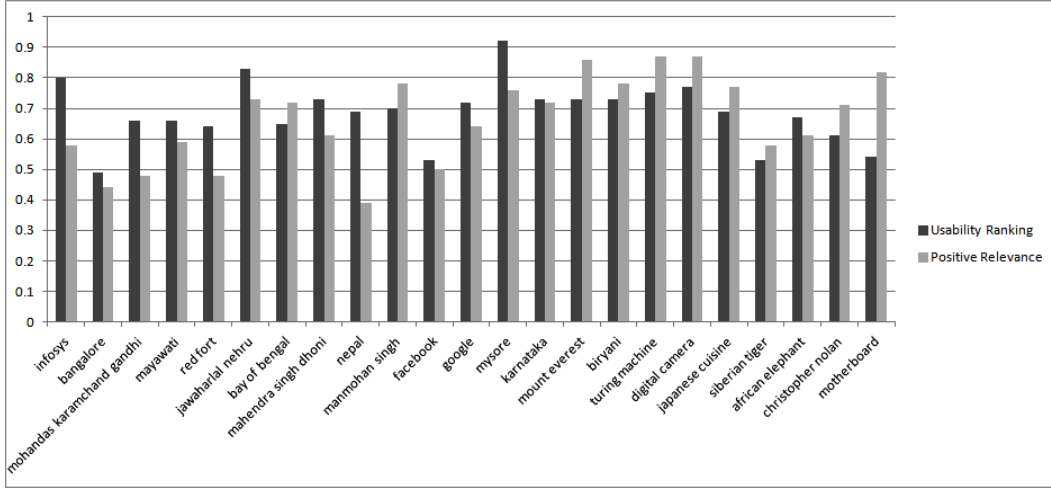


Figure 5.7: Consolidated view of the p_q scores for 23 queries “common” to the two proposed approaches. The X-axis represents the queries, and the Y-axis represents the p_q scores.

While the Usability Ranking approach outperforms the Positive Relevance approach for 13 of these queries, the Positive Relevance approach outperforms the Usability Ranking approach for the remaining 10 queries. However, over these 23 queries, the average difference between the p_q scores of both these approaches is only 0.11. This difference in their performance seems marginal. This indicates that both these approaches show promise in addressing the problem of assigning object-attribute relationships in social space datasets.

Moreover, the overlap between the top-10 attributes generated by the two approaches seems to be low. For each of the 23 queries described above, say q , we measured the Jaccard Coefficient of the attribute-sets of the two approaches as $J(q) = \frac{|attrs_U(q) \cap attrs_P(q)|}{|attrs_U(q) \cup attrs_P(q)|}$, where $attrs_U(q)$ is the set of top-10 attributes for q using the Usability Ranking approach, while $attrs_P(q)$

is the set of top-10 attributes for q using the Positive Relevance approach. Figure 5.8 shows the Jaccard Coefficients computed as above for the various queries.

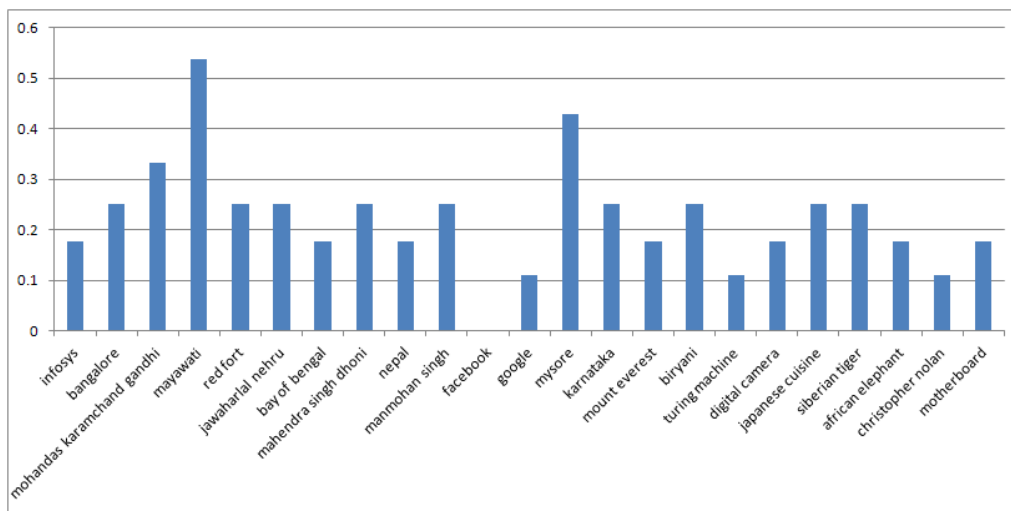


Figure 5.8: Illustration of the similarity (in terms of the Jaccard Coefficient) between the top-10 attributes generated by the two proposed approaches for the “common” queries. The X-axis represents the queries, and the Y-axis represents the Jaccard Coefficients.

These Jaccard Coefficients range from a minimum of 0 to a maximum of 0.54, the average being 0.22, indicating that the attribute-sets generated by the two approaches are largely independent of one another. Both our approaches could be used in conjunction with each other for detecting a wider range of attributes in social spaces.

Even though we have performed our experiments using Wikipedia as the underlying social space, the techniques presented in this chapter are general enough to be applicable to other forms of social spaces as well. In Wikipedia, we consider an article-section as a cognitive context within which concepts

co-occur, and an article as the socio-cognitive process (SCP). Likewise, if we consider a social image sharing platform like Flickr⁴ as the underlying social space, the cognitive contexts would be represented by the tag cloud, the description and each of the comments associated with an image, while the SCP would be represented by the page containing the image along with its tags, its description and comments of other users. The tags co-occurring in the context of an image, the terms co-occurring in the context of the description, and the terms co-occurring within each of the comments would constitute the concepts of this space. Another example of a social space would be a social content sharing platform like Facebook.⁵ Here, the cognitive contexts would be represented by the post itself as well as each of the comments associated with it, while the SCP would be collectively represented by the post and its comments. The terms co-occurring within the post and the terms co-occurring within each of the comments would constitute the concepts of this space.

⁴<http://www.flickr.com/>

⁵<http://www.facebook.com/>

6

Concluding Remarks

The idea behind this thesis was to address the problem of extracting semantics from online information spaces. However, information spaces around us are not of a uniform nature. We therefore classified information spaces into repository spaces and social spaces, based on the differences between them in terms of the nature of cognitive processes governing content creation in them, and the nature of social interactions between such cognitive processes. The cognitive processes lead to local world-views that actors seem to possess

in these spaces. Further, we cast the problem of mining semantics in information spaces as inferring the global world-view held by the population at large.

In order to mine such semantics, we relied on analyzing the co-occurrence patterns of concepts that are of interest. We have argued that co-occurrence analysis is a manifestation of the principles of Ordinary Language Philosophy (OLP) and Hebbian Theory. The idea behind OLP is that the meaning of a term is defined by the way it is used along with other terms within a given cognitive context. On the other hand, Hebbian Theory asserts that the human brain tends to form associations between concepts based on the way the concepts co-occur across a large number of cognitive contexts. Corroborating the principles of OLP, this suggests that the human brain tends to perceive the meaning or identity of a concept based on the way it associates that concept with other concepts. Given that semantics are global world-views of a population, which come to be embedded in information spaces through the cognitive activities of human actors, it can now be argued that co-occurrence analysis, which emulates OLP and Hebbian Theory, is a suitable methodology for mining such semantics. We extend this argument and assert that co-occurrence analysis not only allows us to identify the meaning of a concept in a given context, but also allows us to identify the type of meaning (i.e. the semantic association) it acquires in relation to other concepts within that cognitive context.

We have studied the problem of mining the global world-view held by the population in a repository space. In particular, the artifacts of interest to us were the linkages between documents, which form the basis of the social

interaction between various cognitive processes within the space. Here, the world-view of interest to us were citations that are deemed to be relevant to a certain topic by the population in the repository space. In other words, we looked to identify citations that were endorsed by co-citations. Our work suggests that mining endorsed citations using the co-occurrences of citations may help in fine-tuning existing methods for resource discovery and relevance ranking in repository spaces.

We have also studied the problem of mining the global world-view held by a population in a social space in terms of semantic associations between concepts. The artifacts of interest to us were noun phrases within occurrence contexts in a social space. Here, the world-view of interest to us were the semantic attributes of an object as held by the population in the social space. We proposed two hypotheses for identifying the attributes of a given concept, based on the co-occurrence patterns of concepts. The results of this work suggest that co-occurrence analysis could be effective in addressing the problem of mining semantic associations between concepts in a social space.

We now discuss some of the limitations of the work presented in this thesis.

6.1 Limitations of our Work

It is pertinent to note that, in our analysis on repository spaces as well as social spaces, we have modeled co-occurrence patterns between *pairs* of artifacts, i.e. as binary relations. However, co-occurrence patterns can be modeled as n -ary relations, in general. Suppose a document A has cited a

set of documents \mathbf{B} . Now, if we model the co-citation of $\{A\} \cup \mathbf{B}$ as a *set* by other documents, then we could explore the endorsement of the citation from A to B collectively, instead of the endorsement of citations between just pairs of documents. Here, the arity of the co-occurrence relation would be $|\{A\} \cup \mathbf{B}|$, instead of 2.

Similarly, in the case of modeling usability patterns of concepts in social spaces also, the co-occurrences could be modeled as n -ary relations. This might help in identifying the attributes of an object more accurately. However, for the sake of simplicity, we have modeled only pairs of co-occurrence patterns in our work. This allows us to demonstrate the idea of mining semantics using co-occurrence analysis in a simple fashion. We now discuss some limitations of our approach pertaining to the specific problems we have addressed in this thesis.

6.1.1 Citation Endorsement

We now describe some of the limitations of our work on endorsed citations in repository spaces. Our idea of citation endorsement relies on the co-citation of documents. This poses a disadvantage for documents that are newly created in the dataset, since such documents are not likely to have been co-cited with any other documents. Thus, endorsed citations cannot be identified from new documents. A document will have to have accrued a threshold number of co-citations, especially with the documents that it has itself cited, before endorsed citations from it can be identified.

In methodological terms too, our approach has some limitations. For

instance, we define a citation on the Web as nepotistic if its source and target documents have the same hostname. This precludes the possibility that a citation within the same hostname could actually be topically relevant. Also, this does not guarantee that citations across hostnames are always genuine. Such a decision could potentially impact the quality of endorsed citations on the Web.

Moreover, in the case of the CiteSeer corpus, we do not address the problem of nepotistic citations at all. In reality, it is possible that a document is cited for reasons other than topical merit. A simple heuristic that could be employed in this case is to disallow citations between documents that have common authors. However, it is possible that an author cites her own paper in order to position the current work or to demonstrate incremental work. Here too, such a citation could actually be topically relevant. The identification of topically relevant citations with high precision and recall is crucial for deriving latent semantics using co-citations in a repository space.

Also, we currently assume that the underlying dataset (i.e. the snapshot of the repository space) is static. We have not addressed any measures in this work for incremental handling of the growth of the repository space, so that new endorsed citations can be identified, and current ones updated, dynamically.

6.1.2 Attribute Detection

We now describe some of the limitations of our work on identifying object-attribute relationships in social spaces. Our idea of object-attribute relation-

ships relies on the co-occurrence patterns of concepts. In our work, we define concepts as named entities extracted from the underlying Wikipedia corpus. Given a Wikipedia article-section, we treat as named entities those phrases that form the titles of the target articles of hyperlinks originating in that article-section. However, it is possible that a Wikipedia article-section links to a topically unrelated article for the purpose of elucidation or explanation. This leads to the dilution of the semantic context of the topic and adds noise to it. For instance, the Wikipedia page on *Barack Obama*¹ links to articles on *J-1 Visa*, *Capital Punishment* and *Lame Ducks*. While these links are not directly relevant to the topic of Barack Obama, they have been used as *references* to explain certain terms that occur on that Wikipedia page.

On the other hand, there could be certain other concepts on a given page, which could actually be topically relevant, but do not have a Wikipedia hyperlink anchored around them. Due to our approach, we end up not considering them as named entities. The identification of topically relevant concepts with high precision and recall is very crucial in building topically focused semantic contexts for a given object. If the semantic context is noisy or does not contain genuinely related concepts, the quality of attribute detection for that object could be compromised.

As described in Chapter 5, we asked a group of volunteers to evaluate the precision of our experiments on attribute detection. However, we have not measured the recall of our approaches. A simple heuristic for measuring recall is to ask the volunteer to enumerate what she thinks are the attributes of a

¹http://en.wikipedia.org/wiki/Barack_Obama. Last accessed 17 February 2013.

given object, and to compare this list with the result set of our experiments.²

Also, as with our work on repository spaces, here too our experiments have been conducted on static snapshots of a social space. We have not addressed any measures in this work for incremental handling of the growth of the social space, so that new object-attribute relationships can be identified dynamically.

6.2 Future Work

In conclusion, we discuss some directions for future work in the area of mining semantics in information spaces, in addition to the directions presented by the limitations of our work as described above.

6.2.1 Short-term Directions for Future Work

Distinctions between Repositories and Social Spaces In Chapter 1, we have discussed the theoretical distinctions between repository spaces and social spaces. However, we have not validated these differences experimentally. In the near future, we intend to design appropriate experiments to test the significance of these differences, as well as the differences in the semantics yielded by repository spaces and social spaces.

Applications of Topical Backbones for Specific Repositories In the near future, we also intend to construct endorsed citation graphs for

²This heuristic assumes that the volunteer is sufficiently knowledgeable on the given topic to be able to enumerate the object-attribute relationships.

specific corpora such as Indian publications in Computer Science. One of the applications envisaged for this is the development of appropriate decision support systems for stakeholders such as funding bodies and review boards related to specific academic domains. Another application envisaged for topical backbones of specific repository spaces is browser add-ons for surfing digital libraries using endorsed citations.

Interpretations of Co-citations In Chapter 4, we have introduced three interpretations of co-citations in a repository space. However, we have studied only one of these interpretations in detail in this thesis. In the near future, we intend to explore the other two interpretations, knowledge aggregation and conditional relevance, as well. We will look to use these interpretations to study their applications in topic classification and link prediction, respectively, in repository spaces.

Experiments with Various Datasets We have conducted our experiments on repository spaces with CiteSeer and a Web crawl as the datasets. As mentioned earlier, we intend to conduct similar experiments on other corpora such as Indian publications in Computer Science, High Energy Physics literature, etc. It would also help to try to construct topical backbones (ECGs) using our approach from a sufficiently “noisy” dataset (i.e. comprised of various broad topics) such as arXiv.org. This could give us an idea as to how the ECGs help in the distillation of topics. Further, assuming that topics are distilled into sets of ECG components, it might be interesting to measure the focus of each component (subtopic) of a set (topic) in terms of an aggregate mea-

sure of the citation-endorsement probabilities of the citations within that component. For our experiments with social spaces, we have used Wikipedia as the dataset. However, it might be interesting to perform the task of object-attribute detection in other social spaces such as blogs and Facebook crawls too.

Attribute-Detection and Topic Modeling In Chapter 5, we argue that the task of detecting object-attribute relationships is not the same as the task of topic modeling. However, it would be useful to experimentally compare our approaches with topic modelers such as LDA [Anthes, 2010; Blei and Lafferty, 2007; Blei et al., 2003; Hofmann, 1999a]. This would give us an idea as to how our approaches fare differently from topic modelers. It would also be useful to compare our approaches for attribute detection with existing approaches for ontology learning and relationship mining, which we have discussed in Chapter 2. The existing approaches for topic modeling, ontology learning and relationship mining could be used as a baseline against which the proposed attribute detection hypotheses could be validated.

6.2.2 Long-term Directions for Future Work

Analytical Querying The overarching goal of this work is an analytical system, which can be used to conduct analytical queries about semantics latent within information spaces. Such a system is envisaged to have the following characteristics: (i) A logical data model for modeling content and explaining latent semantics, and (ii) An expressive

query model for being able to conduct queries over the latent semantics. The work done in this thesis, along with the work done by [Mani, 2011; Mutalikdesai and Srinivasa, 2006; Rachakonda et al., 2012], is a step towards modeling latent semantics. In the future, we intend to explore this problem further, and look to design appropriate data models and query models based on co-occurrence patterns of concepts, for online analytical processing over information spaces.

Cognitive Models for Latent Semantics In this thesis, we have modeled information spaces as well as semantics from the perspective of human cognition. We have also positioned co-occurrence analysis as mimicking human cognition. However, this argument needs to be explored further so that the ideas of semantics, cognition and co-occurrence analysis can be seamlessly integrated into an analytical model. The work done by [Rachakonda et al., 2012] is a step in this direction. They propose a 3-layer analytical model based on the episodic as well as the semantic natures of declarative human memory. In the future, we intend to model the semantics extraction problems presented in this thesis in terms of the 3-layer analytical model.

Complete Problem Domain of Semantics in Information Spaces In Chapter 1, we have mapped out the problem domain for semantics extraction in information spaces (see Table 1.1). In this thesis, we have addressed only a subset of this problem domain. In the future, we would like to address the following as well: (i) co-occurrence based semantics extraction in repositories using concepts such as named entities and

topics, and (ii) co-occurrence based semantics extraction in social spaces using social interactions such as comments, “likes”, trust/distrust, etc. One of our objectives is to adapt the semantics extraction algorithms for the entire problem domain to the 3-layer analytical model mentioned above.

Further Classification of Information Spaces In Chapter 1, we have classified information spaces into two categories: repository spaces and social spaces. However, the possibility of classifying repositories and social spaces further into smaller subclasses needs to be investigated. This might help in addressing the issue of semantics extraction from information spaces in a more focused manner. Consider repository spaces. Scientific literature and Web hypertext are two prominent examples of repository spaces. However, there are subtle differences between them. For example, in scientific literature, a paper typically cannot change substantially once published – neither in terms of the content, nor in terms of citations. On the other hand, web pages can be modified to any extent, at any point of time. This also means that citations in scientific literature cannot be cyclic, due to the temporal nature of scientific literature corpora. Citations on the Web, however, can be cyclic. Based on such differences, it might be possible to classify repository spaces further. Similarly, consider social spaces. A wiki corpus incorporates interactions not only in the form of back-and-forth exchange of ideas within a page, but also in the form of citations between pages. A social network such as Facebook, on the other hand,

typically does not have inter-linkages between posts, and incorporates social interactions using comment/votes only within the context of a post. Based on such differences, it might be possible to classify social spaces also further. However, formal definition and modeling of such smaller subclasses is required in order to understand how semantics get embedded differently in each of them.



NP-Hardness of Determinability

We prove that finding the set Q of attributes of x , which has the highest “determinability” for x (i.e. the highest usability of x with respect to Q), is NP-hard, meaning: neither can it be solved by a polynomial time algorithm, nor can a claimed solution for it be verified by a polynomial time algorithm [[Cormen et al., 1990](#)]. We first show that the determinability problem, labeled *DETERMIN* for short, is intractable by showing that every problem in NP can be reduced to the determinability problem in polynomial

time. We then show that $DETERMIN \notin \text{NP}$.

A.1 Intractability of Determinability

We show that every problem in NP can be reduced to the determinability problem by showing that the circuit-satisfiability problem can be reduced to it in polynomial time. We first consider an “easier” version of the $DETERMIN$ problem, labeled $DETERMIN_k$, where the objective is to find the set Q of attributes of x such that $\rho(x|Q) = k$ for some $k \in [0, 1]$. We show that $DETERMIN_k$ is intractable by showing polynomial time reducibility of the circuit satisfiability problem (labeled $CIRCUIT_SAT$ for short) to $DETERMIN_k$. The $CIRCUIT_SAT$ problem is known to be NP-complete [Cook, 1971].

Theorem 1. *The $DETERMIN_k$ problem is intractable.*

Proof. In the $CIRCUIT_SAT$ problem, a given circuit consists of a set L of n input lines. For some input sets $Q \in 2^L$, the circuit produces an output of 1, while for the other sets of inputs, it produces an output of 0. That is, there exists a function f such that

$$f : 2^L \rightarrow \{0, 1\} \tag{A.1}$$

We now map a given circuit in $CIRCUIT_SAT$ to a given entity x whose set of attributes is to be identified in $DETERMIN_k$. For the given circuit, we assume $|L| = |N(x)|$. Given an entity x , we bijectively map the set L of input lines to $N(x)$ using a function g – i.e. $g : L \rightarrow N(x)$ and $g^{-1} :$

$N(x) \rightarrow L$. Here, the function g has a time complexity of $O(|L|)$ (as does g^{-1}). Similarly, we bijectively map the set of outputs $\{0, 1\}$ to $\{-k, k\}$ using a function h – i.e. $h : \{0, 1\} \rightarrow \{-k, k\}$ and $h^{-1} : \{-k, k\} \rightarrow \{0, 1\}$. Here, h and h^{-1} have a constant time complexity. Therefore, using the functions g and h , the function f now be reduced, in polynomial time, to

$$f : 2^{N(x)} \rightarrow \{-k, k\} \quad (\text{A.2})$$

The $DETERMIN_k$ problem produces an output of k when the corresponding problem of $CIRCUIT_SAT$ produces an output of 1 for a given combination of inputs, while the $DETERMIN_k$ problem produces an output not equal to k when the corresponding $CIRCUIT_SAT$ problem produces an output of 0 for a given combination of inputs.

□

This shows that the problem of finding a set Q of attributes of x , such that the determinability of Q for x is some *pre-defined* value k , is intractable. Hence, the problem of finding the set Q of attributes of x , such that Q has the *highest* determinability for x , is also intractable.

A.2 Determinability is not in NP

Lemma 1. *Given a set A with determinability k for a given entity x , the determinability of supersets of A for x does not increase or decrease monotonically.*

Proof. The various combinations of entities from $N(x)$ can be represented

as a poset $(2^{N(x)} \subseteq)$.¹ The Hasse diagram for the poset $(2^{N(x)} \subseteq)$ for an example set $N(x) = \{a, b, c\}$ is shown in Figure A.1.

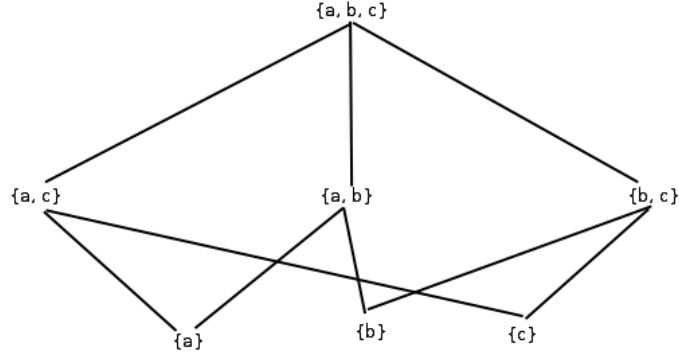


Figure A.1: Illustration of the poset $(2^{N(x)} \subseteq)$

In Figure A.1, as we move upwards from the individual singleton sets towards their supersets, we assert that the determinability of the supersets for an entity x neither increases monotonically nor decreases monotonically. This can be demonstrated using Figures A.2 and A.3 as follows.

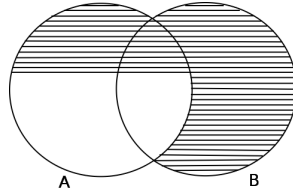


Figure A.2: Illustration of *increase* in determinability in a superset of A

As defined earlier, given a set of entities $A \subseteq N(x)$, the determinability of A for x can be given by:

$$\det(A, x) = \rho(x|A) = \frac{|A_{\perp} \cap C^x|}{|A_{\perp}|} \quad (\text{A.3})$$

¹In fact, $(2^{N(x)} \subseteq)$ is a lattice, since every pair of elements in $2^{N(x)}$ can be shown to have a supremum (least upper bound) and an infimum (greatest lower bound).

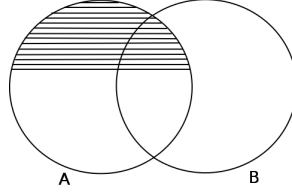


Figure A.3: Illustration of *decrease* in determinability in a superset of B

where A_{\perp} represents the focus of A , i.e. the set of occurrence contexts in which all the elements of A occur, while C^x represents the set of contexts in which concept x occurs.

We prove the above lemma by considering the determinability of the set A and the determinability of the set $A \cup B$, which is a superset of A .

Consider Figures A.2 and A.3. In these figures, consider set A . Let the striped area of A , denoted by s_A , represent $A_{\perp} \cap C^x$, while the area covered by the set A as a whole represents A_{\perp} . Now, the area covered by A is composed of the striped area, s_A , and the plain area, p_A . Thus, Equation A.3 can be re-written to characterize the determinability of A for x as:

$$\det(A, x) = \frac{s_A}{s_A + p_A} \cong \frac{s_A}{p_A} \quad (\text{A.4})$$

Now consider the superset $A \cup B$. Here too, the striped area of $A \cup B$, denoted by $s_{A \cup B}$, represents $(A \cup B)_{\perp} \cap C^x$, while the area covered by the set $A \cup B$ as a whole represents $(A \cup B)_{\perp}$. The area covered by $A \cup B$ is composed of the striped area, $s_{A \cup B}$, and the plain area, $p_{A \cup B}$. Thus, Equation A.3 can be re-written to characterize the determinability of $A \cup B$ for x as:

$$\det((A \cup B), x) = \frac{s_{A \cup B}}{s_{A \cup B} + p_{A \cup B}} \cong \frac{s_{A \cup B}}{p_{A \cup B}} \quad (\text{A.5})$$

In Figure A.2, we see that $\frac{s_A}{p_A} < \frac{s_{A \cup B}}{p_{A \cup B}}$. That is, the determinability of a superset of A , for x , has *increased*. However, in Figure A.3, we see that $\frac{s_A}{p_A} > \frac{s_{A \cup B}}{p_{A \cup B}}$. That is, the determinability of a superset of A , for x , has *decreased*.

□

Therefore, the determinability of supersets of a given set, for a given concept, does not increase or decrease monotonically.

Theorem 2. *The DETERMIN problem is not in NP.*

Proof. If the determinabilities of supersets of a given set, for a given concept x , would *increase* monotonically, then given a set $Q \in 2^{N(x)}$ with determinability k , it would have been possible to verify in polynomial time whether Q has the highest determinability by checking whether $Q = N(x)$. On the other hand, if the determinabilities of supersets of the given set for x would *decrease* monotonically, then it would have been possible to verify in polynomial time whether Q has the highest determinability by checking whether Q is a singleton set.

However, according to Lemma 1, there is an absence of monotonicity in the increase or decrease of determinability of supersets of a given set of attributes. In such a case, we would need to compute the determinability of every $Q \in 2^{N(x)}$ and compare it with k , in order to establish whether Q has the highest determinability. This is a combinatorial computation for verifying the maximal determinability of Q . Hence, $DETERMIN \notin NP$.

□

From Theorem 1, it follows that every problem in NP can be reduced to the determinability problem in polynomial time. Also, according to Theorem 2, this problem is not in NP. Therefore, the determinability problem is strictly NP-hard.

B

List of Related Publications

1. Mandar R. Mutalikdesai and Srinath Srinivasa. Co-citations as Citation Endorsements and Co-links as Link Endorsements. *Journal of Information Science*, volume 36, issue 3, pages 383–400, 2010.
2. Mandar R. Mutalikdesai, Srinath Srinivasa and Viswanath Gangavaram. EndorSeer : An Add-on for Browsing Digital Libraries with “Endorsed” Citations. *Proceedings of the 15th International Conference on Man-*

agement of Data (COMAD 2009), Mysore, India, December 2009.

3. Mandar R. Mutalikdesai and Srinath Srinivasa. On the Semantics of Co-citation, Co-linking and Co-tagging. *Proceedings of the IBM Workshop on Collaborative Academic Research Exchange (I-CARE)*, New Delhi, India, October 2009.
4. Siddhartha Reddy K., Srinath Srinivasa and Mandar R. Mutalikdesai. Measures of “Ignorance” on the Web. *Proceedings of the 13th International Conference on Management of Data (COMAD 2006)*, New Delhi, India, December 2006.
5. Mandar R. Mutalikdesai and Srinath Srinivasa. An Online Analytical Processing Framework for Large Hypertext Collections. *Proceedings of the VLDB 2006 PhD Workshop* (co-located with the 32nd International Conference on Very Large Data Bases), Seoul, Republic of Korea, September 2006.

Bibliography

- H. Abdi. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage*, pages 1–7, 2007.
- S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the 12th International Conference on World Wide Web*, pages 280–290. ACM, 2003.
- L. Adamic. Zipf, power-laws, and Pareto: A ranking tutorial. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>. Last accessed 23 May 2012.
- L. Adamic and B. Huberman. Zipf’s law and the Internet. *Glottometrics*, 2002.
- R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM, 1993.

- M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Proceedings of the SDM 2006 Workshop on Link Analysis, Counter-terrorism and Security*, 2006.
- M. Andrews and G. Vigliocco. The hidden markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, 2(1):101–113, 2010.
- G. Anthes. Topic models vs. unstructured data. *Communications of the ACM*, 53:16–18, December 2010. ISSN 0001-0782. doi: <http://doi.acm.org/10.1145/1859204.1859210>. URL <http://doi.acm.org/10.1145/1859204.1859210>.
- E. Barbu. Acquisition of common sense knowledge for basic level concepts. *Recent Advances in Natural Language Processing*, 1, 2009.
- D. Benz, A. Hotho, and G. Stumme. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference*, 2010.
- D. Benz, C. Körner, A. Hotho, G. Stumme, and M. Strohmaier. One tag to bind them all: Measuring term abstractness in social metadata. *The Semantic Web: Research and Applications*, pages 360–374, 2011.
- S. Bilke and C. Peterson. Topological properties of citation and metabolic networks. *Physical Review E*, 64(3), 2001.
- E. Blanco, H. Cankaya, and D. Moldovan. Commonsense knowledge extrac-

- tion using concepts properties. *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches*, pages 341–353, 2012.
- D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.
- D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer networks*, 33(1):309–320, 2000.
- M. Brunzel. The XTREEM methods for ontology learning from Web documents. In *Proceedings of the Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 3–26. IOS Press, 2008.
- M. Brunzel and M. Spiliopoulou. Acquiring semantic sibling associations from Web documents. *International Journal of Data Warehousing and Mining*, 3(4):83–98, 2007.
- J. Budd. Citations and knowledge claims: Sociology of knowledge as a case in point. *Journal of Information Science*, 25(4):265, 1999.
- P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12, 2005.

- Y. Cao, C. Cao, L. Zang, Y. Zhu, S. Wang, and D. Wang. Acquiring common-sense knowledge about properties of concepts from text. In *Proceedings of the Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 155–159. IEEE, 2008.
- X. Carreras and L. Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics, 2005.
- M. Cegłowski, A. Coburn, and J. Cuadrado. Semantic search of unstructured data using contextual network graphs. http://lists.knowledgesearch.org/papers/Contextual_Network_Graphs.pdf, 2003. Last accessed 23 May 2012.
- S. Chakrabarti. *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann, 2003.
- Z. Chang, Y. Xu, S. Zhang, W. Hu, Z. Li, X. Wang, L. Yu, and H. DuanMu. Relationship mining among the entities associated with GPCRs. In *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, pages 292–295. IEEE, 2009.
- P. Chen. The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.
- T. Chklovski. *Using analogy to acquire commonsense knowledge from human contributors*. PhD thesis, MIT, 2003.

- M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 659–664, 2005.
- S. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151–158. ACM, 1971.
- B. Coppola, A. Moschitti, and D. Pighin. Generalized framework for syntax-based relation mining. In *International Conference on Data Mining*, pages 153–162. IEEE, 2008.
- T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, 1990.
- C. Cottrill, E. Rogers, and T. Mills. Co-citation analysis of the scientific literature of innovation research traditions. *Science Communication*, 11(2):181, 1989.
- I. Dagan, F. Pereira, and L. Lee. Similarity-based estimation of word co-occurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 272–278, 1994.
- I. Dagan, L. Lee, and F. Pereira. Similarity-based models of word co-occurrence probabilities. *Machine Learning*, 34(1):43–69, 1999. URL <http://www.springerlink.com/index/T1T876515PQG5457.pdf>.
- B. Davison. Topical locality in the Web. In *Proceedings of the 23rd Annual*

- International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 272–279. ACM, 2000.
- J. de Ridder, J. Kool, A. Uren, J. Bot, L. Wessels, and M. Reinders. Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. *Bioinformatics*, 23(13):i133–i141, July 2007. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/23/13/i133>.
- J. Dean and M. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11-16):1467–1479, 1999.
- S. Deerwester, S. Dumais, G. Furnas, and T. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Sciences*, 41:391–407, 1990.
- M. Efron. Cultural orientation: Classifying subjective documents by co-citation analysis. In *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*, 2004a.
- M. Efron. The liberal media and right-wing conspiracies: Using co-citation information to estimate political orientation in Web documents. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 390–398. ACM, 2004b.
- U. Essen and V. Steinbiss. Co-occurrence smoothing for stochastic language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 161–164. IEEE, 1992.

- R. Falk and M. Bar-Hillel. Probabilistic dependence between events. *The Two-Year College Mathematics Journal*, 14(3):240–247, 1983.
- S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer. On local estimations of PageRank: A mean field approach. *Internet Mathematics*, 4(2-3):245–266, 2007.
- H. Foundalis. Phaeaco: A cognitive architecture inspired by bongard’s problems, 2006.
- E. Garfield. Citation indexing, historio-bibliography and the sociology of science biography. *Current Contents*, 15, 1971.
- L. Getoor and B. Taskar. *Introduction to statistical relational learning*. The MIT Press, 2007.
- I. Glöckner, S. Hartrumpf, and H. Helbig. Automatic knowledge acquisition by semantic analysis and assimilation of textual information. In *Proceedings of Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS)*, 2006.
- B. Harish, D. Guru, S. Manjunath, and B. Kiranagi. A symbolic approach for text classification based on dissimilarity measure. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pages 104–108, Allahabad, India, 2010. ACM. ISBN 978-1-4503-0408-5. doi: <http://doi.acm.org/10.1145/1963564.1963581>. URL <http://doi.acm.org/10.1145/1963564.1963581>.
- S. Hattori, H. Ohshima, S. Oyama, and K. Tanaka. Mining the Web for

- hyponymy relations based on property inheritance. *Progress in WWW Research and Development*, pages 99–110, 2008.
- T. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.
- D. Hebb. *The organization of behavior*. Wiley & Sons, New York, 1949.
- P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. *Technical Report, Stanford University*, 2006.
- T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999a.
- T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1999b.
- J. Hou and Y. Zhang. Effectively finding relevant Web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):940–951, 2003.
- J. Iria, L. Xia, and Z. Zhang. WIT: Web people search disambiguation using random walks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 480–483, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=1621474.1621581>.

- G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543. ACM, 2002.
- S. Jin, H. Lin, and S. Su. Query expansion based on folksonomy tag co-occurrence analysis. In *Proceedings of the International Conference on Granular Computing*, pages 300–305. IEEE, 2009. ISBN 978-1-4244-4830-2. doi: 10.1109/GRC.2009.5255110. URL <http://dx.doi.org/10.1109/GRC.2009.5255110>.
- G. Kasneci, S. Elbassuoni, and G. Weikum. MING: Mining informative entity relationship subgraphs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 1653–1656, 2009.
- J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. The Web as a graph. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 1–10, 2000.
- K. Lamberts and R. Goldstone. *The handbook of cognition*. SAGE, 2005. ISBN 9780761972778. URL http://books.google.com/books?id=g_mVFLk4kNOC.
- M. Lange, N. Sreenivasulu, A. Stephanik, and U. Scholz. Data relationship mining in life science databases. In *International Workshop on Integrative Bioinformatics in Complex Metabolic Networks*, pages 9–11, 2005.

- R. Larson. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the Annual Meeting of the American Society for Information Science*, volume 33, pages 71–78, 1996.
- R. Lempel and S. Moran. SALSA: A stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160, 2001.
- D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.
- L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.
- K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208, 1996.
- K. Lund, C. Burgess, and R. Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, volume 17, pages 660–665, 1995.
- D. MacKay and L. Peto. A hierarchical dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1995.
- S. Mani. Computing signatures for semantic contexts in online social sp-

- aces. Master's thesis, International Institute of Information Technology, Bangalore, India, 2011.
- L. Màrquez, X. Carreras, K. Litkowski, and S. Stevenson. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159, 2008.
- S. Maskery, Y. Zhang, R. Jordan, H. Hu, J. Hooke, C. Shriver, and M. Lieberman. Co-occurrence analysis for discovery of novel breast cancer pathology patterns. *IEEE Transactions on Information Technology in Biomedicine*, 10(3):497–503, 2006. ISSN 1089-7771. doi: 10.1109/TITB.2005.863863.
- S. McDonald and M. Ramscar. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–616, 2001.
- P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15, 2007.
- E. Minkov, W. Cohen, and A. Ng. Contextual search and name disambiguation in email using graphs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM, 2006. ISBN 1-59593-369-7. doi: <http://doi.acm.org/10.1145/1148170.1148179>. URL <http://doi.acm.org/10.1145/1148170.1148179>.

- G. Moise. *Focused co-citation: Improving the retrieval of related pages on the Web*. PhD thesis, University of Alberta, 2003.
- T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.
- M. Mutalikdesai and S. Srinivasa. An online analytical processing framework for large hypertext collections. In *Proceedings of the VLDB 2006 PhD Workshop*, 2006.
- M. Mutalikdesai and S. Srinivasa. Co-citations as citation endorsements and co-links as link endorsements. *Journal of Information Science*, 36(3):383, 2010. ISSN 0165-5515.
- Z. Nikoloski, N. Deo, and L. Kucera. Degree-correlation of a scale-free random graph process. In *Proceedings of the European Conference on Combinatorics, Graph Theory and Applications*, 2005.
- L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. *Technical Report, Stanford University*, 1999.
- B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008. ISSN 1554-0669. doi: 10.1561/15000000011. URL <http://dl.acm.org/citation.cfm?id=1454711.1454712>.
- B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification

- using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics, 2002. doi: <http://dx.doi.org/10.3115/1118693.1118704>. URL <http://dx.doi.org/10.3115/1118693.1118704>.
- H. Park and M. Thelwall. Hyperlink analyses of the World Wide Web: A review. *Journal of Computer-Mediated Communication*, 8(4), 2003.
- M. Patel, J. Bullinaria, and J. Levy. Extracting semantic representations from large text corpora. *Proceedings of the 4th Neural Computation and Psychology Workshop*, pages 199–212, 1998. URL <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Extracting+Semantic+Representations+from+Large+Text+Corpora#0>.
- J. Pitkow and P. Pirolli. Life, death, and lawfulness on the electronic frontier. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 383–390. ACM, 1997.
- M. Poesio and A. Almuhareb. Extracting concept descriptions from the Web: The importance of attributes and values. In *Proceedings of the Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 29–44. IOS Press, 2008.
- S. Pradhan, K. Hacioglu, W. Ward, J. Martin, and D. Jurafsky. Semantic role parsing: Adding semantic structure to unstructured text. In *Proceedings of the International Conference on Data Mining*. IEEE, 2003.
- A. Preston. *Analytic philosophy: The history of an illusion*. Continuum, 2007.

- R. Quillian. Semantic memory. *Semantic Information Processing*, 1968.
- A. Rachakonda and S. Srinivasa. Incremental aggregation of latent semantics using a graph-based energy model. In *Proceedings of the Symposium on String Processing and Information Retrieval*, Glasgow, UK, 2006.
- A. Rachakonda and S. Srinivasa. Vector-based ranking techniques for identifying the topical anchors of a context. In *Proceedings of the International Conference on Management of Data (COMAD)*, 2009a.
- A. Rachakonda and S. Srinivasa. Finding the topical anchors of a context using lexical co-occurrence data. In *Proceedings of the 18th Conference on Information and Knowledge Management*, pages 1741–1744. ACM, 2009b.
- A. Rachakonda, S. Srinivasa, S. Kulkarni, and M. Srinivasan. Cognitive models for mining latent semantics. *Technical Report, International Institute of Information Technology, Bangalore*, 2012.
- D. Rao and D. Yarowsky. Typed graph models for learning latent attributes from names. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.
- D. Rao, M. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
- R. Rapp. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th Inter-*

- national Conference on Computational Linguistics*, pages 1–7, 2002. doi: <http://dx.doi.org/10.3115/1072228.1072235>. URL <http://dx.doi.org/10.3115/1072228.1072235>.
- P. Reddy and M. Kitsuregawa. Inferring Web communities through relaxed co-citation and dense bipartite graphs. In *Proceedings of the Data Base Engineering Workshop*, 2001.
- S. Reddy, S. Srinivasa, and M. Mutalikdesai. Measures of “ignorance” on the Web. In *Proceedings of the International Conference on Management of Data (COMAD)*, 2006.
- S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.
- F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, and M. Romacker. An environment for relation mining over richly annotated corpora: The case of GENIA. *BMC bioinformatics*, 7(3), 2006.
- D. Rohde, L. Gonnerman, and D. Plaut. An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Science*, 2004.
- B. Russell and J. Slater. *The philosophy of logical atomism and other essays, 1914-19*. Collected Papers of Bertrand Russell. Allen & Unwin, 1986. ISBN 9780049200746. URL <http://books.google.com/books?id=Fpb-9eRqjsYC>.

- M. Sahlgren. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.
- A. Saka and M. Igami. Mapping modern science using co-citation analysis. In *Proceedings of the 11th International Conference on Information Visualization*, pages 453–458. IEEE, 2007.
- P. Sarkar, D. Chakrabarti, and M. Jordan. Nonparametric link prediction in dynamic networks. *arXiv Preprint – arXiv:1206.6394*, 2012.
- C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. *Data Science and Classification*, 2006.
- E. Schwarzkopf, D. Heckmann, D. Dengler, and A. Kröner. Mining the structure of tag spaces for user modeling. In *Online Proceedings of the Workshop on Data Mining for User Modeling*, page 63, 2007.
- F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002. ISSN 0360-0300. doi: <http://doi.acm.org/10.1145/505282.505283>. URL <http://doi.acm.org/10.1145/505282.505283>.
- S. Shapiro. Introduction to SNePS 3. *Conceptual Structures: Logical, Linguistic, and Computational Issues*, 2000.
- H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.

- H. Small. Macro-level changes in the structure of co-citation clusters: 1983–1989. *Scientometrics*, 26(1):5–20, 1993.
- H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu. Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 322–335. ACM, 2009.
- W. Song and S. Park. A novel document clustering model based on latent semantic analysis. In *Proceedings of the Third International Conference on Semantics, Knowledge and Grid*, pages 539–542, 2007. ISBN 0-7695-3007-9. doi: 10.1109/SKG.2007.169. URL <http://portal.acm.org/citation.cfm?id=1338447.1338958>.
- S. Soundarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st International Conference on World Wide Web*, pages 607–608. ACM, 2012.
- M. Steyvers and T. Griffiths. *Probabilistic topic models*. Lawrence Erlbaum Associates, 2007. ISBN 1410615340. URL <http://www.worldcat.org/isbn/1410615340>.
- N. Sundaresan and J. Yi. Mining the Web for relations. *Computer Networks*, 33(1-6):699–711, 2000.
- E. Terra and C. Clarke. Frequency estimates for statistical word similarity measures. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Hu-*

- man Language Technology*, 2003. doi: 10.3115/1073445.1073477. URL <http://portal.acm.org/citation.cfm?doid=1073445.1073477>.
- M. Thelwall. *Link analysis: An information science approach*. Emerald Group Pub Ltd, 2004.
- M. Thelwall and D. Wilkinson. Finding similar academic web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 40(3):515–526, 2004.
- P. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the European Conference on Machine Learning*, 2001.
- C. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, 1977.
- L. Vaughan. Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, 57(9):1178–1193, 2006.
- L. Vaughan and J. You. Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3):611–628, 2006.
- L. Vaughan, M. Kipp, and Y. Gao. Why are websites co-linked? The case of Canadian universities. *Scientometrics*, 72(1):81–92, 2007.
- A. Veling and P. van der Weerd. Conceptual grouping in word co-occurrence networks. In *Proceedings of the Sixteenth International Joint Conference*

- on Artificial Intelligence*, pages 694–701, 1999. ISBN 1-55860-613-0. URL <http://portal.acm.org/citation.cfm?id=646307.687429>.
- C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 203–212. ACM, 2010.
- D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1100–1108. ACM, 2011.
- M. Wettler and R. Rapp. Computation of word associations based on the occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora*, pages 84–93, 1993.
- H. White and B. Griffith. Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981.
- H. White and K. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186, 1989. ISSN 0066-4200.
- H. White and K. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355, 1998.

- L. Wittgenstein. *Philosophical investigations*. 1953. URL <http://books.google.com/books?id=XN9yyyhYMDoC>.
- D. Zhao. Going beyond counting first authors in author co-citation analysis. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*. Wiley Online Library, 2005.
- J. Zhu, Z. Nie, J. Wen, B. Zhang, and W. Ma. 2D conditional random fields for Web information extraction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1044–1051. ACM, 2005.
- J. Zhu, Z. Nie, J. Wen, B. Zhang, and W. Ma. Simultaneous record detection and attribute labeling in Web data extraction. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 494–503. ACM, 2006.