AUTOMATIC HYPERTEXT CONSTRUCTION

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
James Allan
January 1995

© James Allan 1995
ALL RIGHTS RESERVED

AUTOMATIC HYPERTEXT CONSTRUCTION

James Allan, Ph.D.
Cornell University 1995

The unprecedented growth of the World Wide Web illustrates the importance of hypertext as a method for organizing the rapidly expanding amount of on-line text. As document collections become larger and more dynamic, however, it is not feasible to construct more than an occasional hypertext manually. This thesis presents entirely automatic methods for gathering documents for a hypertext, linking them, and annotating those connections with a description of the type or nature of the link.

The problem of automatically collecting related documents is addressed in Chapter 2, where robust Information Retrieval methods are applied to form high-quality links between documents. A local context check identifies links where ambiguous vocabulary erroneously suggests a relationship. Dynamic part retrieval is employed to select the portions of documents which are most related, allowing parts to be linked when it is more appropriate to link subtopics than entire documents.

Chapter 3 presents a taxonomy of hypertext link types and defines the following three classes of links: “pattern-matching” links can be found using simple string-matching methods, “manual” links require substantial application of natural language understanding methods (which are currently beyond the state of the art), and “automatic” links are those which can be found using the methods of this thesis.

Chapter 4 begins the work of automatic link typing by describing two novel graphical techniques for visualizing the relationship between two or more documents. “Uniform” visuals display the relationship between documents or document parts without regard to their relative sizes, whereas “varying” visuals include information about sizes and locations. Both methods highlight relationships between documents and motivate the automatic techniques of Chapter 5.

Chapter 5, thus, demonstrates automatic methods for identifying the relationships depicted in the visualizations. Using an approach based upon graph simplification, this method automatically identifies revision, summary, expansion, equivalence, comparison, contrast, tangential, and aggregate links.

Chapter 6 discusses an informal evaluation of the link typing. Though somewhat inconclusive, the evaluation demonstrates that automatic document linking performs

well, but also indicates that much work remains to be done toward understanding automatic link typing.

Biographical Sketch

Shortly after his debut on June 19, 1961, James Allan enjoyed a full year run in Grinnell, Iowa, a small town admidst the cornfields. Though his performance was a smashing success, his agents (affectionately known as "Mom" and "Dad") felt the opportunities were better on the east coast, so they began what turned into a 17-year run in Carlisle, Pennsylvania.

The small mid-western audiences still held some fascination, though, so James eventually headed back out to Grinnell for a production which culminated in his winning the coveted A.B. degree in 1983. James then returned to Carlisle for a 5-year association with Dickinson College. The critically acclaimed production of *Autocat* was one of the most enjoyable and rewarding experiences of his career.

In 1988, James chose to move north to Ithaca and begin a new life at Cornell University. His work there earned him an M.S. in 1992 and the Ph.D. in January of 1995. James has accepted a post doctoral research fellowship with the Center for Intelligent Information Retrieval at the University of Massachusetts in Amerst. He intends to remain in academics, but continues to debate an appropriate balance between teaching and research. And theatre.

To Mom and Dad
role models, advisors, parents
friends

Acknowledgements

Being a graduate student has been a great deal of fun. Yes, the dread of qualifiers, the tedium of some classes, the occasional desire to seriously harm an officemate, the usually dreadful grading sessions, and the nail-pulling agony of finishing a thesis—all that unpleasantness made graduate student life less than idyllic. But the pleasure of passing the Q's (at last), the loosely structured lifestyle, the opportunity to learn almost anything, the striking beauty of Ithaca's environs, the joys of friendship, the comraderie of "AI" on Fridays, and the thrill of putting one past the goalie—all those more than made up for the relatively minor unpleasantness.

I am grateful to my advisor, Gerard Salton, for all he taught me about Information Retrieval (IR) research: what it is, what it should be, and (perhaps most importantly) what it should not be. His unique combination of a relaxed atmosphere and a driven research program were exactly the environment I needed for my studies. I feel privileged to have worked under him.

I am also tremendously indebted to Chris Buckley, the architect of the Smart system software which formed the basis of most of my research. Without his help I would still be trying to understand the system. My fellow IR student, Amit Singhal, also enriched my time here with numerous lengthy and informative debates about the field. My research was supported in part by the National Science Foundation under grants IRI 89-15847 and IRI 93-00124.

The people in the department have been especially kind. The cleaning crew, the support staff, and "adm" have all made my life brighter and easier, but I would like to single out Jean Mills, Anne Gockel, Debbie Smith, Becky Personius, and Jan Batzer for special recognition. Among the faculty, Bruce Donald, Keith Marzullo, Dexter Kozen, and David Gries have encouraged me in ways they probably little realize. Devika Subramanian and Daniela Rus have been good friends as well as good colleagues.

My minor field of study was theatrical directing under the auspices of David Feldshuh. David is one of the most fascinating men I have had the opportunity to work with and I thank him profusely for all he taught me, but especially for the lessons in creativity.

My Cornell experience was greatly enriched by the time I spent with my good friends Russell Brown, Jim Jennings, Judith Underwood, Paul Jackson, William Rucklidge, Sam Weber, and the many others whom I have no doubt forgotten to mention.

I am, however, especially grateful to some very special friends: to Brad and Cindy Glade for their generosity and kindness, to Cindy Robinson for caring and for reminding me to laugh, to Paul Stodghill for enduring me and for reminding me to relax, to Alessandro Panconesi for his humor and for reminding me to be compassionate, and to Mark Hayden for the camaraderie and for reminding me to think.

Finally, I thank my family for their love and support: Mom for her concern, Dad for his advice and his example, Karen for her faith, and Susan for her enthusiasm (even if she did beat me in the race). Such a wonderful family is a rarity and I have been blessed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Motivation

The amount of textual information accessible via computer networks is growing at an unprecedented rate. Without some means of organization, the abundance of information becomes merely an overload of data with little or no meaning. Some efforts have been made to address this problem by creating repositories of data and providing search engines for scanning the data,[Kah91] or by decentralizing that approach somewhat and providing a means for searching through descriptions of databases.[Kah91] Information Filtering services and research[FD92] are gaining renewed interest as the amount of on-line information and the number of on-line information consumers both grow.[YGM93,Har92a] Research into tools for browsing a collection of texts has resulted in several new and possibly useful techniques.[CKP93,CU94]

However, measured by the amount of accessible information, the number of users, and the rate of its growth, by far the most successful information organizing tool on the network is the World Wide Web. In 1992, the Web accounted for about 500 megabytes of traffic on the Internet (0.001% of total traffic); in the first quarter of 1993, it accounted for 5 gigabytes of traffic (0.03% of the total); over 10 gigabytes (5%) on the single day May 1, 1994; and in October of 1994, over 2.1 terabytes (10%).[Gra94] As of November 8, 1994, the Web is estimated to consist of 6.3 gigabytes of data in almost a million documents.[Mau94,Fou94]

The World Wide Web (WWW) is a loosely organized collection of documents forming a hypertext[1] scattered throughout the Internet. Any user on the computer network may create a Web document and publicize its presence. Other users may then access the document and possibly include references to it in their own documents. Although some documents or sets of documents are disconnected from the rest of the hypertext, almost any of the hundreds of thousands of documents may be accessed

[1] The term *hypermedia* is usually used when the linked data can include information other than text—*e.g.*, images, audio—as it can on the WWW. This work is concerned only with text, however, so we will use the term *hypertext* throughout.

directly or indirectly from any of the others.

As the Web and other hypertexts grow and become more dynamic, it becomes less and less likely that the hypertext links can continue to be created solely by manual means. For example, it is unimaginable that the 400 megabytes of data in roughly 130,000 articles of network news articles created *per week*[Rei94] could be meaningfully linked by any non-automatic means.

Motivated by the explosion of hypertext growth and use, this thesis explores questions related to the *automatic* construction of hypertext. It uses document similarity functions adapted from the field of Information Retrieval to show how related documents can be found automatically. The major contribution of the thesis, however, is the next step, where a description of the *nature* of the documents' relationship is automatically determined. This description provides a "type" for the link, a necessary component of any complex hypertext system.

1.1 "Lost in hyperspace"

The concept of "hypertext" was originated by Vannevar Bush in 1945[Bus45] when he described a "memex" system for organizing and retrieving information. He envisioned "...a device in which an individual stores all his books, records, and communications and which is mechanized so that it may be consulted with exceeding speed and flexibility." He continued by describing how a person might call up two documents and enter a link between the two of them. At a later point, when one of the documents is displayed, the link may be followed to retrieve the other effortlessly.

Bush's vision may have been somewhat unrealistic at the time, but it is eerily on target from a modern perspective, and even by the late 1960's, it was already a reality. The Brown University Hypertext Editing System was not only a text editing system, but also "a reading machine on which to browse and query written materials having complex structure"[CGN+69] (one of the authors of that system was Ted Nelson who had coined the term "hypertext" a few years earlier). The "oN-Line System" (NLS) developed by Engelbart and others at the Augmented Human Intellect Research Center at Stanford Research Institute represented text in an outline-like hierarchy and allowed arbitrary links between the parts of the text.[EE68]

In the midst of bold statements that hypertext would eventually replace the printed word, there were already rumblings of potential pitfalls:

> Since it is rather easy to get lost in a complicated hypertext, we plan to look into displaying its graph structure in a variety of ways. A parts graph may be drawn for the simple case, but how does one display several hundred cross links in one area?[CGN+69]

By 1974, Nelson was expressing concern for the potential complexity of the hypertext networks—"It could also get you good and lost"[Nel74]—though he optimistically suggested that a simple stack of previously visited documents (a backtracking mechanism) would eliminate the problem.

Nelson was correct that a history tracking approach is useful, and most modern hypertext systems include such a feature. However, such a simple navigation aid is far from sufficient to avoid disorientation in the hypertext. Numerous other approaches have been suggested, among them: commands to go directly to known locations,[AMY88] graphical maps and browsers,[Noi93,Fri88] even making all links "invisible" to reduce clutter.[IB90]

This work, however, focusses on one particular approach to addressing the navigation problem: link typing.

1.2 Importance of link typing

A link in a hypertext is, at its most basic level, a connection between two units of text. Without any further information describing the link, it serves merely as a starting point for jumping to an unknown location. A few systems (such as the World Wide Web) do not support link types *per se*,[2] forcing the hypertext author to include link annotations in the text itself (*e.g.*"for a definition of *hypertext* click here").

A *type* is an attribute assigned to a link. (The word "type" in this context bears only some similarity to its use in programming languages—as in the "type" of a variable—and the two notions should not be confused. For example, a link could be assigned multiple, quite contradictory types, an disallowed situation in most programming languages.) The link attribute gives some idea of the effect of following the link, thereby yielding a more powerful hypertext.[You90] The importance of link types is regularly mentioned in the literature:

- Parunak talks about how typed links can be used selectively to simplify the overall topology. "If links are *classified* by different types, the topology induced by links of any one type may be much simpler than the overall topology of the entire system."[Par89]
- Conklin argues that one solution to the "lost in space" problem "...is to apply standard database search and query techniques to locating the node or nodes which the user is seeking. This is usually done by using boolean operations to apply some combination of keyword search, full string search, and logical predicates on other attributes (such as author, time of creation, type, etc.) of

[2] The Web has links which behave differently—*e.g.*, which invoke `ftp` or which use a `gopher` protocol—but the logical nature of the link is not part of the link itself.

nodes or links."[Con87] That is, the link type can be used to select only links which interest the user at the moment.

- Lucarella describes how link types can be useful when an inference-style information retrieval system is being used within a hypertext.[Luc90] (This approach is even more valuable if real-valued weights or "belief values" are included with the link type—something which the techniques of this thesis can easily provide.)
- Marshall and Shipman claim that "... [links] can be used to articulate specific semantics for interconnection. ...These links make structure explicit, and are thus useful in writing, argumentation, problem structuring, or capturing semi-formal knowledge representations of a domain."[MI93]
- A hypertext browsing system may choose to behave differently when different types of links are selected by the user. For example, a "glossary" link might cause a pop-up window to be displayed, while a "view related documents" might replace the current document with a summary of several documents.

Almost all commercial and research hypertext systems provide at least a minimal notion of a link type. Some examples include:

- Engelbart's NLS system[EE68] had simple types based mostly upon document structure—e.g, link to list tail, list head, sub-heading, *etc.*.
- Xerox PARC's Notecard allowed the user to provide a label for every link which indicated the link's purpose.[Hal88] (Irler and Barbieri call these "labelled" rather than typed links.[IB90])
- KMS supports two basic link types: structural and annotation.[AMY88]

1.2.1 Types of link types

Despite the importance of link types, it is surprising that there is so little agreement about what link types should exist. Most hypertext systems appear to add link types in an *ad hoc* manner, trying to add all the ones that occur to the programmers.

At least two researchers, however, have attempted to create a list of all hypertext link types that can occur. At first, Trigg's list of the 80 link types[Tri83] seems a bit presumptuous, but he does not claim it is exhaustive. Instead, he argues that the disadvantage of limiting the number to 80 types is out-weighed by the special processing that might be possible given a known set of link types. Others agree that some advantage may accrue from forcing a hypertext author to choose from a limited set of types: "Link types provide a local view of the roles of individual links but give little indication of how the set to which the links belong is structured. With the variance of link types from system to system, particularly where link types are user-defined, they may not even be consistently interpreted or generally understood."[You90] Using a pre-defined (albeit not all-inclusive) set of link types would help with interpretation.

Parunak draws from the discipline of Discourse Analysis to develop a taxonomy of possible hypertext links.[Par91] It is not clear that Parunak's taxonomy will prove to be universal, but its nature is inherently extensible: newly defined link types can usually be located within the taxonomy. Chapter 3 describes Parunak's list in more detail.

1.2.2 Typed links—a panacea?

Type annotations on links are undoubtedly important and useful for hypertext navigation. However, they do not completely solve the "lost in hyperspace" problem. van Dam rightly claims that "...in a sense hypertext gives us a **goto**, and a **goto**, as we all know, produces spaghetti."[vD88] In that view, a typed link is just a special type of **goto** statement—and "attaching names to arbitrary transfers of control would do little to aid understanding."[You90]

Despite their limitations, the discussion above makes it clear that link types are useful and there are tremendous benefits to be gained if we can determine the link types automatically.

1.3 Manual vs. automatic

Significant research has been done in the area of automatically creating hypertext from linear text, however almost without exception the methods have relied upon pre-existing mark-up for the text (*e.g.*, SGML annotations), or have used a pre-constructed knowledge base to derive the links. The former condition is limiting since so much text exists (and continues to be created) without annotated structure. The latter is limiting because knowledge bases are very difficult to build for unrestricted subject areas.[Sal91]

Bernstein's hypertext "apprentice"[Ber90] is one of a very few systems which use techniques similar to those which will be presented in this thesis. The test-bed for his work was two very small pre-existing hypertexts (each fewer than 200 hypertext "pages") which he attempted to augment by finding additional links. Although Bernstein was moderately successful, he felt the possibilities of automatic linking (using his techniques) were limited and he stopped at that point. His negative conclusions may have been a side effect of using a less sophisticated text analysis tool than that used in this work.

Stotts and Furuta's χTed system can automatically detect structure—but only if the structure clues have been pre-specified in a grammar.[SF90] Wilson's Guide system builds complex legal hypertexts, but relies upon fairly standardized vocabulary in the legal domain.[Wil90]

Dynamic hypertexts such as that proposed by Shibata and Katsumoto[SK93] necessarily provide automatic hypertext links (they are created on demand rather than in advance). However, their system requires an extensive knowledge base to be usable in a general setting.

The methods described in this thesis work in the presence or absence of pre-existing structural mark-up—they can use the mark-up to improve performance, but do not require it. The methods work irrespective of the subject matter of the material in the document collection—it may be specific to any topic, or it may span numerous topics. Finally, these methods do not require any human intervention (though minor manual adjustments can be useful for improving effectiveness slightly).

1.4 Thesis summary

This thesis, then, explores methods for automatically determining the type of a link between two documents or document passages—and for automatically locating the link before typing it. Although there has been substantial research related to building hypertext links and displaying them, this work represents the first significant and successful efforts at creating and typing such links on arbitrary collections of text without user intervention.

Chapter 2 reviews recent Information Retrieval work in combining local context information with global similarities to create a more robust measure of document and query similarity. The chapter continues by reviewing a technique for selecting passages from a document which are believed to answer an information need better than the entire document. The resulting "part retrieval" has been shown to increase the number of significant documents retrieved while simultaneously causing a dramatic reduction in the amount of retrieved text.

Chapter 3 draws upon work from the field of Discourse Analysis to present classes of link types which are useful in a hypertext. The chapter discusses which of those classes can be automatically recognized using simple pattern matching techniques, which can be found automatically using the techniques of this thesis, and which are currently beyond the state of the art for automatic detection.

The major contributions of this thesis are in Chapters 4 and 5. The first presents methods for visualizing and describing the relationship between documents or document passages. New graphical methods are demonstrated which are useful for illustrating such relationships.

Chapter 5 presents algorithms which analyze the pattern of relationships between the documents' component subparts and thereby permit the automatic detection of several relationships: subtopic, related topic, condensed or expanded treatment, extracted passage, revised document, and so on. The resulting relationships between documents can be used to construct a hypertext automatically and to provide some clues to alleviate the navigation problem.

Chapter 6 presents the results of an informal evaluation of the reasonableness of the automatically generated links. The evaluation is made by presenting a set of automatically-generated links to users and asking for their rating of several aspects of the linked documents' relationship. The evaluation does not conclusively support the link typing of Chapter 5, but neither does it refute the approach.

1.5 Experimental setting

All experiments were done using the Smart information retrieval system. Smart was developed over the past 30 years under the auspices of Gerard Salton, primarily at Cornell University. The current version of the system, Version 11, was designed and written primarily by Chris Buckley.

For these experiments, a collection of legal documents from the 1988 and 1989 Federal Register (470Mb from the TREC collection) is used.[Har92b] This collection includes 46,315 documents ranging in size from 94 bytes to over 2.6Mb (the average document size is 10.2Kb). Examples are also taken from the 26,000 article Funk and Wagnalls encyclopedia,[FW79] and from a small collection of Computer Science Technical Reports.

Experiments were run on any of several Sun SPARC workstations.

Chapter 2

Document Linking

The Information Retrieval techniques of clustering and classification both provide natural means for automatically generating groups of related documents. When the documents in a collection are grouped by a sufficiently robust measure of similarity, those groups will provide natural anchors for hypertext navigation.

This chapter presents a high-quality, robust method for automatically finding relationships between documents in a collection. We also show that for larger documents, it is necessary to violate the integrity of the document and find relationships between parts of documents rather than their entirety. The texts—whole or partial documents—which are related to a given document form the desired source and destination of a hypertext link. That link can then be typed as will be described in Chapter 5.

We begin by reviewing the model used for retrieval and then show how retrieval works in practice. Next, a reasonably new global/local restriction method is described which improves the likelihood that matched documents are related. Finally, the issue of passage handling is addressed.

2.1 Vector space model

The well-known vector space model provides the framework for document comparisons throughout this work. In that model, each query and each document in a collection is represented by a *vector*: pairs of concepts and non-negative weights signifying the importance of each concept within that document. Once documents or queries are converted to vectors, they can be compared by measuring the angle between the vectors. If two vectors lie near one another, the assumption is that their corresponding documents or queries are similar.[Sal75]

2.1.1 Term weighting

For small collections, the index concepts could be manually assigned, but when large numbers of documents are being processed automatically, it is standard to use the terms of the documents as the concepts—and automatic techniques are just as effective as manual approaches.[Sal68]

All of the documents in the collection are broken into their constituent words. The words are then reduced to their root form, so that *act*, *actor*, *acting*, and so on are treated as the same concept. The resulting set of t unique concepts is used to represent each document in the collection by assigning a weight to each concept for each document.

In a high-quality system, the term weight is typically based upon three components:

tf (term frequency): the frequency of a concept's occurrence within a document: the more often a concept is used, the more likely it is that the concept is important within that document,and the more heavily weighted that term should be.

idf (inverse document frequency): is related to the number of documents in which the term occurs. A term which occurs in a small number of documents is useful for distinguishing one document from another, whereas a term which occurs in all documents has no value in that regard. So a good term weight is inversely proportional to the document frequency of the term.

norm (normalization): normalization of term weights is used to prevent very long documents—where terms occur frequently and are heavily weighted by the **tf** factor above—from dominating much smaller documents.

A typical non-normalized weight for term T_k in document D_i is:

$$v_{ik} = \underbrace{tf_{ik}}_{\text{“tf”}} \cdot \underbrace{\log(N/n_k)}_{\text{“idf”}} \tag{2.1}$$

where N is the number of documents in the collection, tf_{ik} is the term frequency of term T_k in document D_i, and n_k is the number of documents in which term T_k occurs at least once. A typical *normalized* weight is created by applying length normalization to yield a vector of length 1.0:

$$\begin{aligned} w_{ik} &= \frac{v_{ik}}{\sqrt{\sum_{j=1}^{t} v_{ij}^2}} \\ &= \frac{v_{ik}}{||D_i||} \end{aligned} \tag{2.2}$$

Note that in both equations 2.1 and 2.2, concepts which do not occur in a document are given a weight of zero; all other weights are positive.

With these equations, non-normalized term weights v_{ik} are directly proportional to the term frequency. If documents D_i and D_j are identical except that term T_k occurs twice as often in D_i as it does in D_j, then $v_{ik} = 2v_{jk}$. On the other hand, normalized term weights w_{ik} are related to the *proportional* term frequency—*i.e.*, to the proportion of vector length contributed by that term. So if term T_k occurs twice as often in D_i as in D_j (and as before, the documents are otherwise the same), then:

$$\begin{aligned} w_{ik} &= \frac{v_{ik}}{\sqrt{\sum_{\ell=1}^{t} v_{i\ell}^2}} \\ &= \frac{2v_{jk}}{\sqrt{(\sum_{\ell=1}^{t} v_{j\ell}^2) + 3v_{jk}^2}} \\ &= \frac{2v_{jk}}{\sqrt{||D_j||^2 + 3v_{jk}^2}} \\ &= w_{jk}\frac{2}{\sqrt{1 + \frac{3v_{jk}^2}{||D_j||^2}}} \\ &= w_{jk}\frac{2}{\sqrt{1 + 3w_{jk}^2}} \end{aligned}$$

So the increase in the term's weight depends upon that term's contribution to the total length of the vector: the more "significant" the term, the less its weight will increase.

It is important to understand the effect of normalization on term weights so that an appropriate weighting scheme can be selected. For example, when the weight should clearly reflect the number of common terms, normalization is inappropriate.

2.1.2 Vector similarity

Given two vectors, $\vec{u}$ and $\vec{v}$, their similarity is measured by their inner product:

$$\begin{aligned} \mathrm{sim}(\vec{u}, \vec{v}) &= \vec{u} \cdot \vec{v} \\ &= \sum_{i=1}^{t} u_i v_i \end{aligned} \qquad (2.3)$$

If the vectors are normalized to have length 1, equation 2.3 is exactly the cosine of the angle between the two vectors. In that case, similarities range from 0.0 for totally orthogonal vectors (no terms in common) to 1.0 for totally identical vectors (all terms in common and in the same proportion).

When the vectors are not normalized (as in equation 2.1), the similarity is:

$$\mathrm{sim}(\vec{u}, \vec{v}) = \cos\theta \cdot ||\vec{v}|| \cdot ||\vec{u}||$$

Here the similarity is the cosine of the angle between the vectors, multiplied by the length of each vector. The similarity can range from zero for orthogonal vectors to arbitrarily large for extremely long documents.

Because non-normalized term weights are proportional to the term frequency, the similarity in that case will be larger when frequently occurring terms are in common. As a result, non-normalized term weights are useful when the similarity should reflect the *number* of terms in common, and normalized weights are useful when the similarity must reflect the *proportionate* use of terms within the two documents.

2.2 Retrieval

To find a group of documents which is related to a chosen document D_i, the set is selected whose vectors lie closest to D_i's vector. The set is ordered by the cosine of the angle between the vectors, so the most similar documents are listed first. Depending upon the needs of the system, either a fixed number of nearby documents can be retrieved, or only those documents whose vectors are within a chosen ϵ from D_i are chosen. Finding those documents can be done in many ways, but an inverted file is the most efficient for large databases.[Sal75]

Table 2.1 shows the result of using this technique to find the 20 documents most highly related to an encyclopedia article discussing “March music.” The most highly similar document is the starting article itself which has a (normalized) similarity of 1.00 (same terms in the same proportion). Many of the listed documents are clearly related through their discussion of music; however, quite a few documents should not have been selected. The non-relevant articles are marked with an “N” in the “Rel?” column.[1] One document, *Rhythm*, is of marginal relevance—the *Musical rhythm* article is clearly relevant, however. The documents marked “X” are cross reference articles that lead the reader to a more useful article: in this case, every document marked “X” in Table 2.1 points to a relevant document also in the table.

[1] “Relevance” is highly subjective and so can vary greatly with each person. In this case, articles are considered relevant if they mention March-style music in some fashion.

Table 2.1: Documents similar to *March (music)*
(N=non-relevant; X=cross reference; ?=marginal relevance)

	Num	Sim	Rel?	Title
	14966	1.00		March (music)
1.	21548	0.43		Sousa, John Philip
2.	14965	0.39	N	March (month)
3.	14969	0.30	N	Marche
4.	16928	0.21	N	Northern Expedition
5.	14313	0.20	N	Long March
6.	12088	0.20	N	Infantile Paralysis, National Foundation for
7.	15520	0.19	X	Meter (music)
8.	14968	0.19	N	March of Dimes Birth Defects Foundation
9.	16287	0.18		Musical Rhythm
10.	7551	0.17		Drum (musical instrument)
11.	9284	0.17	N	Fraser, Malcolm
12.	14967	0.17	N	March, Fredric
13.	21404	0.14	X	Snare Drum
14.	16282	0.14		Music, Western
15.	19582	0.14	?	Rhythm
16.	12520	0.14		Jazz
17.	12494	0.14	N	Japanese Music
18.	21938	0.14	X	Syncopation
19.	14970	0.13	N	March to the Sea
20.	22038	0.13	X	Tambourine

2.3 Global-local restrictions

The incorrectly chosen documents of the previous section arise because of ambiguous vocabulary usage: *march* can be a type of music, a month of the year, the name of person or organization, or the name of a geographical region. This problem has been the source of substantial research in the two decades since the introduction of the vector space model and is receiving attention more recently also. Often-proposed solutions to this problem appeal to various natural language processing (NLP) or artificial intelligence research areas: determine the part of speech,[XBC94] isolate meaning using a thesaurus,[Voo93,San94] parse and attempt to place the documents in case frames,[CC92] and so on.

Those techniques, however, tend to have difficulty if the subject area is not constrained or if there is inadequate time to train the algorithms.[CC92] Research within the past 5 years has focussed on larger texts, where it is fortunately possible to fall back on simpler techniques. Wittgenstein's "use theory" of meaning states that the meaning of a word is determined completely by how the word is used.[Wit53] This theory suggests that it should be possible to distinguish the meanings of *march* by examining the contexts in which the word occurs.

To that end, if two documents have a sufficiently high similarity, the documents are decomposed into smaller pieces—usually sentences, though paragraphs, sections, groups of sentences, or even phrases are possible—and each piece of one document is compared to each piece of the other. If there is no pair of pieces with similar enough context, then there is no common usage of vocabulary between the two documents.

The context of two pieces is usually compared by considering three measures:

1. The overall similarity between the vectors corresponding to the pieces. Because we are interested in the use of words, the similarity is usually non-normalized—*i.e.*, the number of words in common is important, not their proportion. Values ranging from 10.0 to 75.0 are frequently used.

2. The actual number of terms in common. The similarity is proportional to this number, but it is often desirable to ensure that at least a minimum number of terms match. To require that a single term does not cause the match (a possibility if the term occurs frequently in the document or vary rarely in the collection), this measure typically must have value 2 or 3.

3. The contribution of the most highly weighted term toward the vector similarity. This information is typically used to prevent a very highly weighted term combined with several content-free words, from passing the other criteria. A value of 90% or 95% is typical.

Table 2.2: *March music* documents, sentence match
N=not relevant, ?=marginal relevance, ·=also in Table 2.1

	Num	Sim	Rel?	Title
	14966	1.00	·	March (music)
1.	21548	0.43	·	Sousa, John Philip
2.	16287	0.18	·	Musical Rhythm
3.	7551	0.17	·	Drum (musical instrument)
4.	16282	0.14	·	Music, Western
5.	12520	0.14	·	Jazz
6.	854	0.13		American Music
7.	8815	0.13		Fife (musical instrument)
8.	21933	0.12	?	Symphony
9.	16284	0.12	?	Musical Form
10.	2067	0.12		Band
11.	19977	0.12	N	Rumba
12.	18625	0.12	N	Popular Music
13.	4973	0.11	N	Chamber Music
14.	21519	0.11	N	Sonata
15.	16306	0.11	N	Mussorgsky, Modest Petrovich
16.	18854	0.11	?	Program Music
17.	9046	0.11	N	Folktales
18.	9045	0.10		Folk Music
19.	12038	0.10	?	Indian Music

Table 2.2 shows the result of applying a sentence restriction to the same "March music" search of Table 2.1. In this case, the sentences were transformed into vectors with non-normalized weights. If there did not exist even a single pair of sentences with a similarity of 70.0, with at least 2 terms in common, and with the most heavily weighted term contributing no more than 95% of the similarity, the document was removed from the list.

Table 2.2 clearly shows the effectiveness of this approach: the only documents from Table 2.1 (marked with "·") are the relevant documents which were not cross reference articles. The 15 documents which failed the sentence restriction were replaced by additional documents (numbers 6 through 19 in Table 2.2). Only 4 of the 14 newly retrieved documents are actually relevant to the query, but such a result is not surprising considering the extremely low similarity of the documents compared to the query—in this collection, any similarity below 0.20 is suspect, and anything below 0.15 is quite weak (the dashed line marks the 0.15 similarity boundary). These new documents pass the local restriction, but their global similarity is too weak for the match to be considered significant.

The effectiveness of this relatively new approach has also been demonstrated repeatedly with objective measurements.[SA93, SAB93, SABS94] Recent work has achieved similar success with a slight variation of this approach which computes a new similarity by combining the global and local similarities (rather than discarding documents which fail to achieve a high enough local similarity).[BSAS94]

2.4 Part retrieval

A document collection such as the Federal Register contains documents with sizes ranging from 94 bytes to over $2\frac{1}{2}$ megabytes (more than 500 pages). In collections with such large documents, it is reasonable to expect a user to become tired of scanning large documents for relevant material. It is desirable, therefore, for a retrieval system to automatically determine the portion of a document which is the best match to the query—or to another document. Such a feature has been rare in retrieval systems until quite recently, when the sizes of on-line documents grew substantially: from short abstracts to full texts of documents.

For any retrieval system to select the preferred portion of a document, the document must be broken into smaller pieces. Many modern documents are annotated with some form of mark-up language (*e.g.*, SGML or TEX) which can often be used to find the logical breaks in the text. For non-annotated parts (sentence breaks appear to be marked only rarely), simple pattern matching suffices. Note, however, that the patterns often cause incorrect sentence breaks—they should be used with that in mind. (Determining paragraph breaks with such an approach is straightforward;

Table 2.3: *March music* documents, retrieving sections
N=not relevant, ?=marginal relevance

	Num	Sim		Title
	14966	1.00		March (music)
1.	21548	0.43		Sousa, John Philip
2.	9046.c7	0.27	N	Folktales
3.	12520.c6	0.22		Jazz
4.	7551.c4	0.19		Drum (musical instrument)
5.	16287	0.18		Musical Rhythm
6.	854.c7	0.17		American Music
7.	16282	0.14		Music, Western
8.	2067.c3	0.13		Band
9.	21933.c7	0.13	?	Symphony
10.	8815	0.13		Fife (musical instrument)
11.	21519.c4	0.13	N	Sonata
12.	16284	0.12	?	Musical Form
13.	19977	0.12	N	Rumba
14.	12038.c5	0.12	?	Indian Music
15.	18625	0.12	N	Popular Music
16.	4973	0.11	N	Chamber Music
17.	16306	0.11	N	Mussorgsky, Modest Petrovich
18.	18854	0.11	?	Program Music
19.	9045	0.10		Folk Music

some work has been done toward identifying section boundaries.[RS94])

In Table 2.3, the “March music” search was changed to allow sections of documents to be retrieved. This change has two noticeable effects on the retrieved documents:

1. A section of a document may have higher similarity to the original document than did the entire document. The most noticeable example of that is the *Folktales* document which moved from 17th to 2nd in ranking (perhaps unfortunately given than the document is not relevant to the original document—the *Marchen*, or fairy tale, is one type of folktale). In some cases, new documents may appear in the “top 20” list because a part of the document is strongly related, but the extraneous material in the document as a whole depresses the overall similarity.

   Other than the *Folktales* article, every case where a section of a relevant document was chosen, it is the best possible section of the document: the origin of *Jazz* in small marching bands, how *drums* are used in marching bands, Sousa's march music as part of the history of *American Music*, mention of marching bands as a type of *band.*

2. The total amount of retrieved text is substantially reduced. The documents listed in Table 2.2 amount to roughly 222-thousand characters of text. The documents and sections listed in Table 2.3 amount to only 135-thousand characters, a 39% reduction in presented text.

   This reduction in presented text is useful for sifting through documents for relevance or non-relevance, or for understanding why a document was retrieved. For example, in the *Indian Music* article, the selected section discusses the rhythm and time-keeping aspects of Indian music, presumably the portion of the document interesting to someone interested in march music.

When the part retrieval is extended to allow paragraphs to be selected—if the paragraph is a better match than its containing section—the results change again. Table 2.4 shows that a paragraph of *American Music* is now ranked 2nd, where the entire document (and also its best-matching section) were both ranked 6th. The total amount of text contained in the retrieved documents and parts is 113-thousand characters, a 16% reduction from when sections were allowed, and a 49% reduction from entire document retrieval.

Again, the part retrieval clearly identifies that portion of the document which is believed most relevant to the query. In every case of Table 2.4 where a paragraph was retrieved in place of an entire document, it is the paragraph most related to marching music. In the cases where relevance is marginal, the selected paragraph is the paragraph which is somewhat relevant. Even in those non-relevant cases, the identified paragraph makes it clear why the document is listed—*e.g.*, paragraph 24 of *Popular Music* discusses the rhythmic elements of rock and roll music in a manner very similar to the discussion of rhythm in *March music.*

Little work has been done toward evaluating the effectiveness of part retrieval. This absence is primarily the result of the lack of an experimental collection which includes relevance judgements for parts of documents. (The lack of such a collection is partly due to the difficulty of deciding upon the "most relevant" passage, but is also due to the relatively new nature of document part research.) Some related work has been done which uses passage-query similarity to adjust the global similarity of documents.[Cal94] Some experiments have suggested that retrieving passages improves precision but has little impact on recall.[Wil94] New approaches using, for example, Hidden Markov Models are also being suggested and claimed feasible.[MS94]

Table 2.4: *March music* documents, retrieving paragraphs

	Num	Sim		Title
	14966	1.00		March (music)
1.	21548.p4	0.50		Sousa, John Philip
2.	854.p25	0.44		American Music
3.	9046.c7	0.27	N	Folktales
4.	12520.p16	0.24		Jazz
5.	7551.p8	0.19		Drum (musical instrument)
6.	18625.p24	0.18	N	Popular Music
7.	16287	0.18		Musical Rhythm
8.	16282	0.14		Music, Western
9.	21519.p5	0.14	N	Sonata
10.	16306.p4	0.14	N	Mussorgsky, Modest Petrovich
11.	21933.p14	0.13	?	Symphony
12.	2067.c3	0.13		Band
13.	8815	0.13		Fife (musical instrument)
14.	16284	0.12	?	Musical Form
15.	12038.p12	0.12	?	Indian Music
16.	19977	0.12	N	Rumba
17.	4973	0.11	N	Chamber Music
18.	18854	0.11	N	Program Music
19.	9045	0.10		Folk Music

2.5 Linking documents

If two documents are believed to be similar to one another, there is a relationship between them. If the two documents are part of a hypertext collection, a link can be created between them. Note that there are several options for selecting which documents or passages to link to the starting (query) document:

1. Create a link to the matching document or to the best passage (or passages) of that document. For example, based upon the information in Table 2.4, a link might be created from *March music* to *John Philip Sousa*, or just to the 4th paragraph of the *Sousa* article.

2. Create a link between the parts of each document which are most similar. The analysis above, for example, used the entire *March music* article. However, the 4th paragraph of *Sousa* is most similar to the *3rd* paragraph of *March music.* So it would be reasonable to make the link between the two paragraphs.

3. Create links from the query document to all sufficiently similar documents (or parts). That would mean creating 7 links from the list in Table 2.4.

4. Create links between *all* of the documents similar to the query document. That expands the 7 links to 28 (or 56 if the links are not symmetrical).

5. Create links between documents even when they fail the local similarity check. Such links would be typed differently, but can occasionally be interesting. In the example, that would incorporate the 8 documents marked `N` in Table 2.1.

Using some set of those choices, it is possible to create a hypertext such as that represented in Figure 2.1. In that figure, 42 documents were linked whenever there was a strong-enough similarity between them. That set of documents was constructed by starting with the 4 articles most similar to the article on *March music*, and then adding up to 8 documents most similar to each of those, and finally up to 4 documents most similar to each of those. (This process could have proceeded until no more documents were linked to the set, but the complexity of such a web is already apparent.)

Chapter 5 will discuss these approaches in more detail and determine the link types to which they correspond.

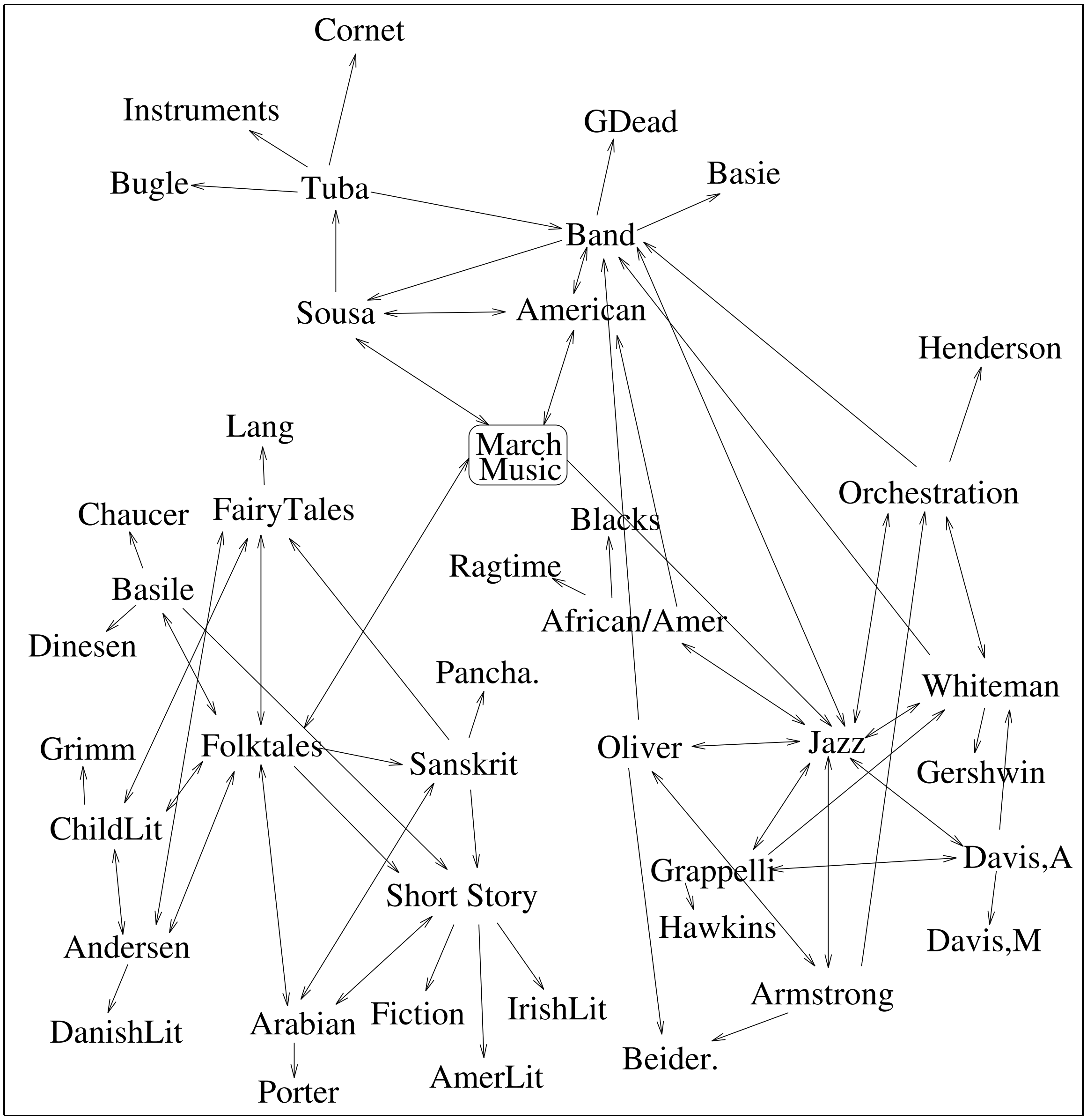


Figure 2.1: Complex hypertext starting from *March music*

Chapter 3

Taxonomy of Hypertext Link Types

A link type is a description of the relationship between the source and the destination of a link. As such, it cannot be determined by considering only the destination document: the document could be an *example* of one concept, and a *counter-example* of another. Nor are link types symmetrical: following a link may lead to an *expanded discussion* of a topic, but returning along the link will clearly not do the same.

In hypertext systems such as the World Wide Web, which do not incorporate link types, a writer entering links into the hypertext is forced to include implicit link types, usually by describing what is at the link's endpoint. (Note that since the description is in the source, the implicit type describes the relationship, not just the destination.) Some examples taken from the WWW show how this is achieved (the link's source is presented in italics):

> If you would like to comment on The Asylum, then use *Quick Takes* or you can send us a *message.*[The94]
>
> The list of CSD reports and their respective prices *is here.*[Ber94]
>
> You may also list all the TRs indexed by this server, sorted either by *author's last name* or by *number.*[Ber94]
>
> Please see the *disclaimers* and other information about our server.[Cor94]
>
> ...important links to the *people* and *places* working to make a customer driven government a reality.[NL94]

Such implicit link types are reasonable in many cases, but become cumbersome when less explicit link types are used or when there are multiple possible destinations from a starting point. Although some of the awkwardness is actually a Human-Computer

Interaction issue,[Sal90] more of it is the result of forcing the link type to be embedded in the running text.

It is preferable, therefore, to include the type of a link with the link itself. The types could be arbitrary labels assigned by the writer,[Hal88] but if the types are from known classes it is possible for hypertext browsing software to take advantage of that information and act in a manner appropriate for that class of link. (This view of links is reminiscent of object-oriented approaches: the "follow link" action is possibly different for each class of link types.)

Chapter 2 demonstrated how Information Retrieval techniques can be used automatically to find related documents and link them together. It further showed that it is possible to select passages automatically from the documents for linking. Chapter 5 will develop automatic methods for describing the relationship between documents or passages more specifically than just "related." In order to do that, however, it is necessary to determine what types of hypertext links can and should exist.

3.1 Taxonomy of link types

Links can be divided into classes in several ways. Trigg provides the major divisions of internal "substance" links and external "commentary" links.[TW87] Each of those is then broken down into sub-classes, which may then be broken into sub-sub-classes, and so on. In all, Trigg provides a set of 80 classes of link types.[Tri83]

Parunak draws upon the field of Discourse Analysis to present "classes of link types that are useful in hypermedia."[Par91] Discourse analysts are interested in understanding the nature of communication with language, and have developed a taxonomy of relationships between "predications" or arbitrary passages of text.[Lon76, Lon83] With that as a basis, Parunak develops his own hierarchy of hypertext link types, including 27 types in 3 broad categories: revision, association, and aggregation.

In this work, we will present an amalgam of known link types and divide them into three major categories based upon whether or not their identification can be achieved automatically. The three categories are Manual, Pattern-matching, and Automatic. Some link types, unfortunately, straddle the boundaries, depending upon document collection being linked—*e.g.*, it is possible to identify some types of links if the subject area is small enough and known in advance, where the link type cannot be recognized in a general setting.

3.1.1 Pattern-matching links

This first broad class of link types are those which can be found easily using simple—or sometimes fairly elaborate—pattern-matching techniques. An obvious example of such a link type is “definition” which can be found by matching words in a document to entries in a dictionary. In almost all cases, these links are from a word or phrase to a small document, and will occur outside of any specific context—*i.e.*, the destination document *may* be the same for the word or phrase, no matter where the word or phrase occurs.

We also group structural links into this class. Structural links are those that represent layout or possibly logical structure of a document. For example, links between chapters or sections, links from a reference to a figure to the figure itself, and links from a bibliographic citation to the cited work, are all structural links. We include this with pattern matching links because they are typically recognized by mark-up codes that are already embedded in the text. Even when a document is not marked up, structure is usually approximated using pattern analysis.[RS94]

The following classes of links which were identified by Parunak and Trigg, fall into the “Pattern-matching” class:

Content links have as their source, words which name a proposition, and as their destination, that proposition. Links to graphics, audio, or other non-textual items are content links, as would be a link from a bibliographic citation to the cited work.

Identification links define or clarify the meaning of a word. Links to a glossary are obvious examples of identification links.

Comment links are like an identification link except that they do not restrict the meaning of the word or phrase in anyway: instead they provide additional, elaborative, information.

Orientation links have as their destination, information about the location, time, or circumstance of the material in another node. Some of these links—*e.g.*, dates, names of companies, names of people, locations, monetary amounts—can be found automatically with simple pattern matching routines.[CU94]

Legal case aggregate links connect the various aspects of a legal case—*e.g.*, fact, issue, rationale, precedent, decision, and so on. Some success has been achieved in automatically isolating some of this information since the legal domain has a fairly well-specified and rigid vocabulary.[Wil90]

Software module aggregate links are actually more like a set of annotations of a program, identifying constants, functions, *etc.*, allowing some form of “re-use” of common features. Since programming languages have a grammar, it is relatively straight-forward to decompose a program automatically into its component pieces and store them in a database. (The difficulty is then determining a specification for the component so that it can later be retrieved for re-use.[CFG91])

Structural links are found by mark-up codes or text patterns that represent the beginning and end of various structural components of a document. In general, structural links are found because a large document is decomposed into several smaller hypertext nodes. The structural links are easily remembered as the document is analyzed.[NLBH88,Stu85]

3.1.2 Manual links

Pattern-matching links are a class which is easy to detect automatically. At the extreme opposite end of the spectrum are "manual" links, those which we are currently unable to locate without human intervention.

Identifying manual links requires text analysis at a level which the Natural Language Understanding community is trying to achieve. They have had some significant success within constrained subject areas, so some "manual" links could be automatically described within those limited domains. Unfortunately, the techniques cannot yet be extended to a general setting, so this class of link types remains inaccessible to automatic approaches.

Manual links include the following classes from Parunak's and Trigg's taxonomies:

Circumstance links connect a document to a second document which describes the circumstances under which the situation occurred.

Argument or **discussion** links collect together the important parts of a debate: the issue, points pro and con, rebuttal, and so on.

Implication links describe relationships such as caused-by, purpose, condition, contrafactual, concession, warning, evidence, and so on.

3.1.3 Automatic links

Between the difficulty of manual links and the ease of pattern-matching links, is the location of "automatic" links. These are links which cannot typically be located trivially using patterns, but which the automatic techniques described in Chapter 5 can identify with marked success.

The automatic links which can be identified are:

Revision links are a very simple class of relationship between texts, including both ancestor and descendent relationships. In the context of computer-edited material where successive versions of the material are archived, either intentionally or as an artifact of the editing system, information which describes relations in some "revision hierarchy" is crucial. Even when the revision occurs over a much greater time—*e.g.*, a revised edition of a textbook—it is useful to know the relationship.

Some revision links are fortunately extremely simple to find and maintain automatically, since they are flagged by the editing system—*e.g.*, version numbers of a file, backup copies of a text. (Those revision links would actually be classified as "pattern matching" links.) When that information is absent, however, a different means must be used to find the relationship.

Summary and **expansion** links are inverses of one another. A summary link type is attached to a link which starts at a discussion of a topic and has as its destination a more condensed discussion of the same topic.

Equivalence links represent strongly-related discussions of the same topic.

Comparison and **contrast** links identify similarities and differences (respectively) between texts.

Tangent links move from one topic to one which is similar, but in an unusual or tangential manner. For example, a link from a document about "Clouds" to one about Georgia O'Keeffe (who painted a mural entitled *Clouds*) would be a tangential link.

Aggregate links are those which group together several related documents. An aggregate link may actually have several destinations, allowing the destination documents to be treated as a whole when desirable.

With the exception of *equivalence* links, found to a limited degree by Bernstein's hypertext "apprentice"[Ber90], no other system had been able to successfully identify these classes of links in a collection of text covering an arbitrary subject area. Chapter 5 explains how these link types can be identified without human intervention. Chapter 4 first presents some novel visualization tools which inspire the approach of Chapter 5.

Chapter 4

Relationship Visualization Techniques

The graph of Figure 2.1 in the previous chapter is one way to represent the relationship between large numbers of documents. It does not, however, give any detailed information about the type of the link. As a first step toward automatically describing the relationship between two or more documents, this chapter presents two novel visualization techniques.

The first technique, *uniform* comparison, treats each significant part or document as an identical unit and displays them and the links between them. This form of visualization is useful for seeing simple relationships, for discovering unusual links, and for some simple structural analysis of a document.

The second technique, *variable* comparison, includes information to reflect the relative size of each part and to make it clear how the part relates to the document containing it. The graphs arising from this approach will be the inspiration for the link typing described in Chapter 5.

4.1 Uniform

Figure 4.1 shows the result of comparing the *March Music* encyclopedia article to the 19 articles most similar to it (Table 2.2 on page 14 shows the result of retrieving those articles). In this graph, each of the 20 documents has been placed on the edge of an oval. If there is a non-zero similarity between two documents, a link is drawn between them. The *March music* article is highlighted in italics and surrounded by a rounded rectangle.

Unless noted otherwise, every graph presented in this chapter was created automatically by the Smart system. For presentation purposes, however, most of them

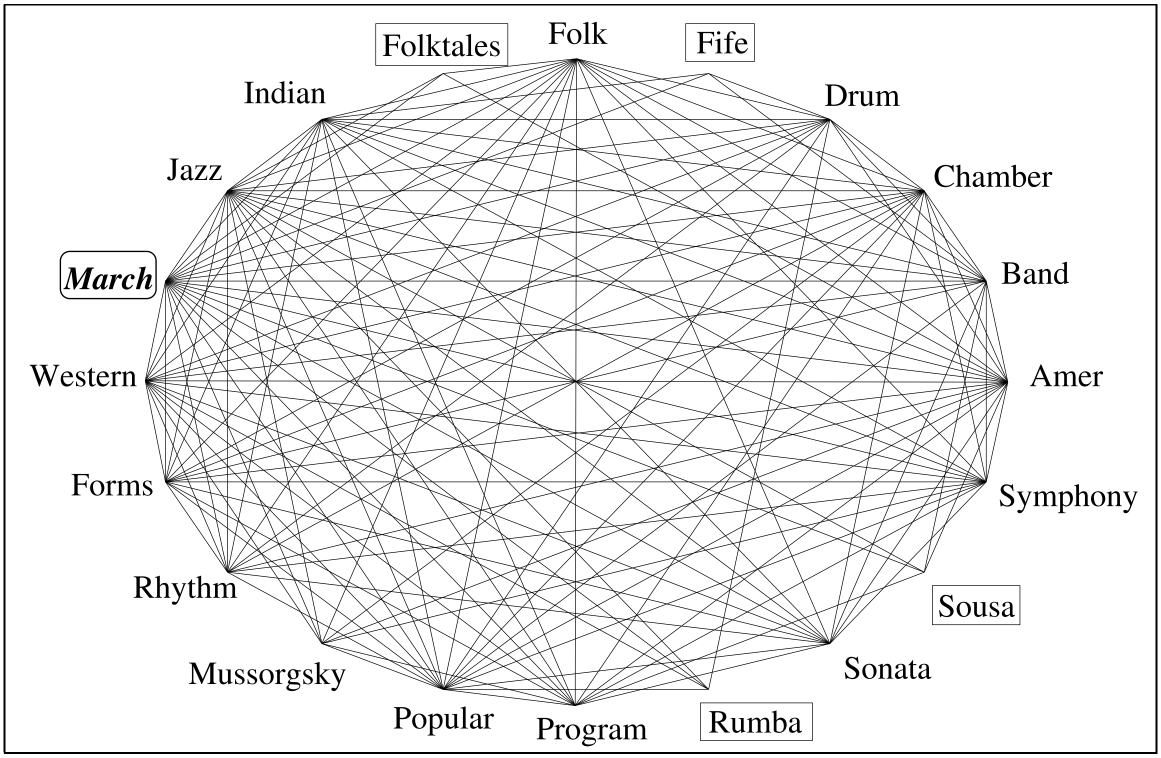


Figure 4.1: Articles similar to *Music*

have been edited by hand. In the case of Figure 4.1, the node annotations (Jazz, Forms, Drum, *etc.*.) were selected and positioned by hand; the titles of articles in the encyclopedia are often long and look aesthetically displeasing in such a complex graph. The links were chosen and graphed automatically.

Even from a graph as dense as Figure 4.1, it is possible to glean some information. With the exception of the four documents with names surrounded by rectangles, every node on the graph has 9 to 19 links—*i.e.*, they are related to half or more of the group. The four highlighted documents are linked to only four or five other documents, clearly suggesting an unusual relationship. They in fact correspond to articles on:

- 8815 *Fife (musical instrument)* – This article discusses the instrument and its use in marching bands. It has very little other discussion, so its limited linking is reasonable.

- 9046 *Folktales* – This article is not relevant to *March music* or even a general discussion of music, so it is not surprising that it has few links.

- 19977 *Rumba* – This article is another non-relevant document, though because it touches upon music it is less inappropriate than the *Folktales* article.

- 21548 *Sousa, John Philip* – This article is strongly tied to *March music* but, like the *Fife* article, contains little other discussion. It's tight focus limits the number of links.

Each of the other documents—those with many links—contains a broader discussion, suggesting that *March music* is not the central theme of these documents.

A variation on the graph of Figure 4.1 is shown in Figure 4.2 where the style of the link is chosen depending on the similarity strength. In this particular case, the line styles are chosen as:

Line Style	Similarity Range	Description
Dotted	0.00–0.20	tenuous
Solid	0.20–0.40	related
Solid, heavy	0.40–0.60	very related
Solid, very heavy	0.60–1.00	strongly related

With similarity information represented in the graph, the somewhat tenuous relationship of three of the four articles listed above is highlighted: they are not related in any strong way to any of the other documents.

But this form of the graph also highlights other interesting features which are useful to the searcher. The set of articles in the graph was chosen by finding the 19 articles most similar to the *March music* article, 14966 (marked in the figure

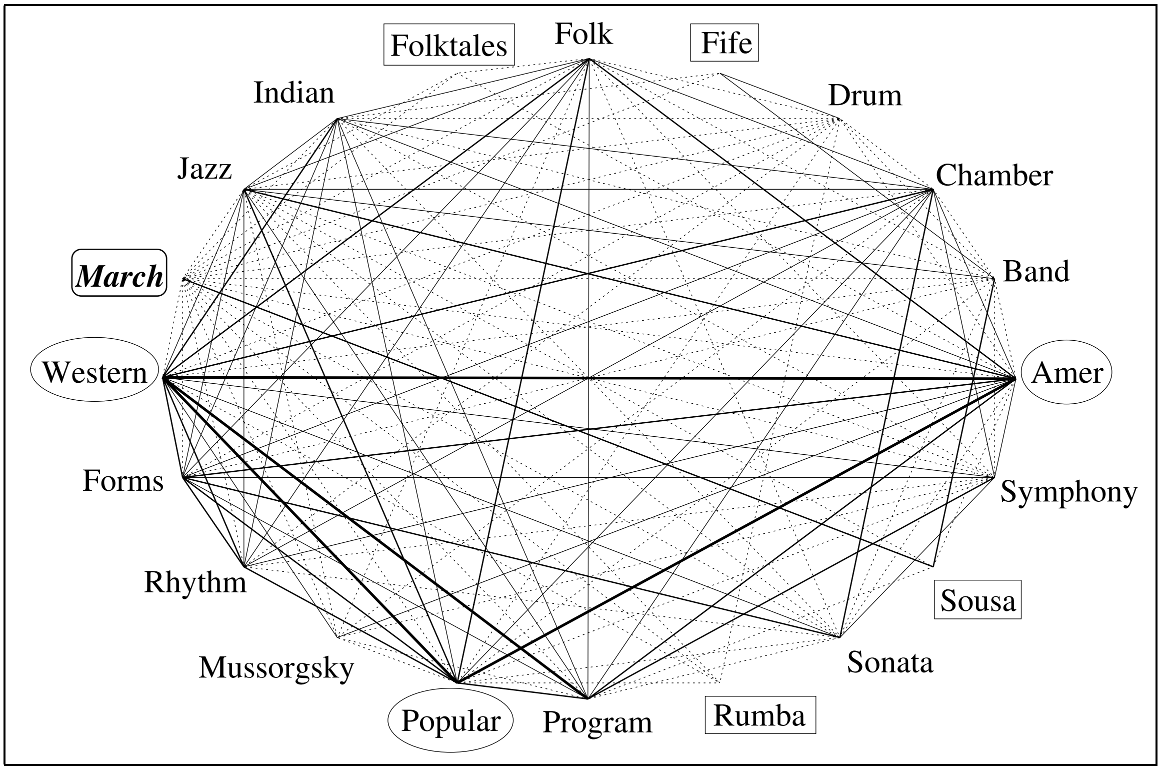


Figure 4.2: Articles similar to *Music*

with italics text and a rounded rectangle). Note that only one article—*John Philip Sousa*—has more than tenuous similarity with the starting article. In fact, the three articles with their names surrounded by ovals—16282 *Western Music*, 18625 *Popular Music*, and 854 *American Music*—are clearly more "central" in that they have strong relationships to more of the 20 articles. Recent work has demonstrated that this theme can be detected and summarized automatically.[SABS94]

It is also possible to find central articles by eliminating the low-similarity links gradually. This process achieves an effect similar to de-emphasizing weak links by making them dotted, but the link reduction process makes it clear which nodes are important. Figure 4.3 shows the progression as the minimum required similarity is slowly raised. (The node corresponding to *March Music* is marked with a rounded rectangle; the rectangles and ovals mark the same documents they did in Figure 4.2.)

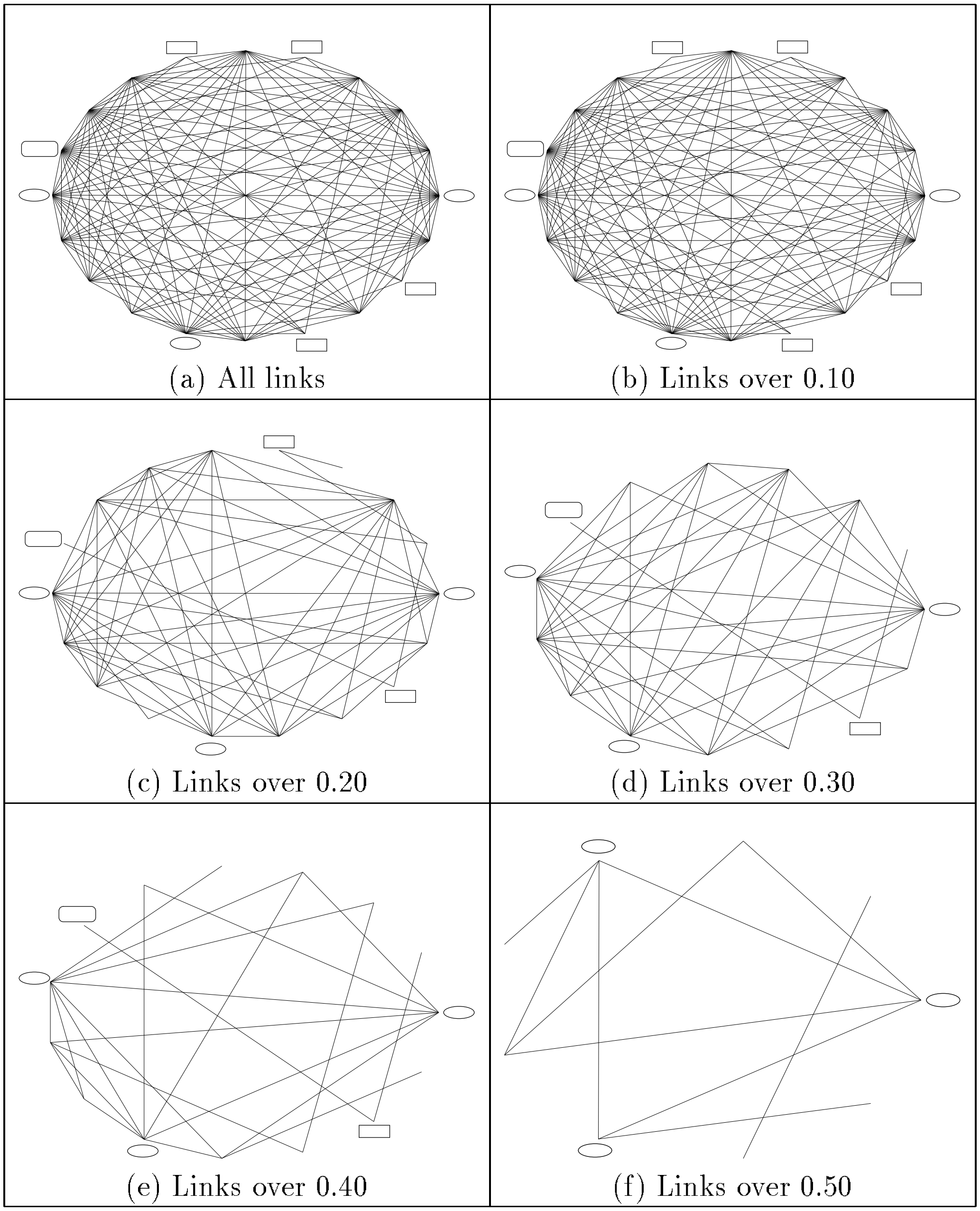
In Figure 4.3a, all articles are included and linked whenever there is a non-zero similarity. In (b), all similarities below 0.20 (tenuous links) are dropped from the figure, and any articles with resulting degree zero are also dropped: as the discussion above suggests might happen, documents 9046 *Folktales* and 19977 *Rumba* are the first two articles dropped. Since neither *Folktales* nor *Rumba* is relevant to the music topic, their loss is an improvement.

When the threshold is raised to 0.30, 3 more documents are dropped: 7551 *Drum (musical instrument)*, 8815 *Fife (musical instrument)*, and 16306 *Mussorgsky, Modest Petrovich*. The first two articles discuss musical instruments used in march music, but do not discuss the more general topic of "music" embodied in this set of documents. The loss of *Mussorgsky* is welcome because it is not relevant to the march music topic.

As the similarity threshold is raised from 0.30, through 0.40 in 4.3d, and to 0.50 in (e), the set of articles still presented becomes more and more tightly related. By the time a threshold of 0.50 has been reached, the starting document (*March music*) has even been eliminated in favor of general discussions of music. Although the group was initiated by a discussion of the *March music* topic, the larger articles on music-related topics swamped the original theme and created a new one.

4.1.1 Uniform visuals with parts

The graphs presented so far in this section depicted relationships between entire documents. But they can also be used to view the relationships between the parts of documents—or between the parts of a single document. Figure 4.4 shows the result of breaking documents 9045 *Folk Music* and 854 *American Music* into their component paragraphs and comparing the paragraphs of each document. The dotted line running left-to-right was added by hand to help separate the paragraphs of *American Music* (above) from those of *Folk Music* (below). Links with a similarity below 0.20 are not displayed and, as before, nodes which have resulting degree zero are suppressed.



(a) All links

(b) Links over 0.10

(c) Links over 0.20

(d) Links over 0.30

(e) Links over 0.40

(f) Links over 0.50

Figure 4.3: Raising similarity threshold

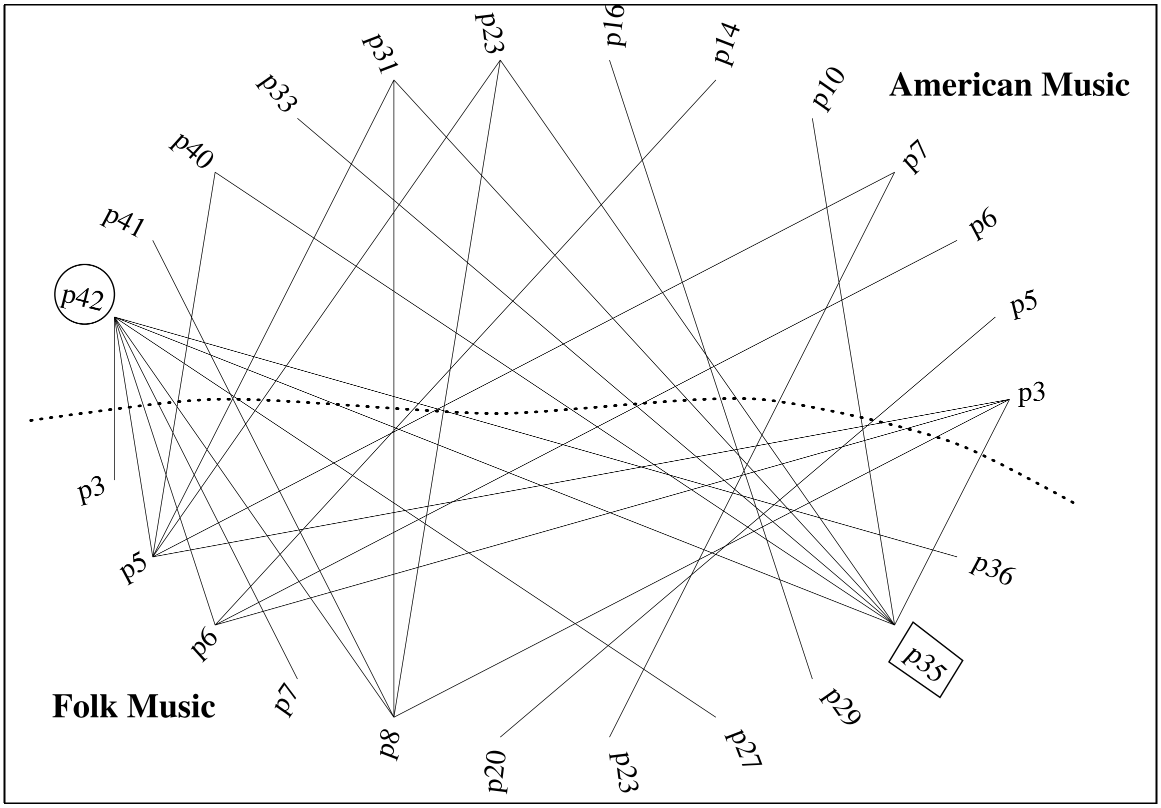


Figure 4.4: *American* and *Folk Music* articles

This comparison of parts is sufficient to show some strong relationship between some paragraphs in one document to large parts of the other. Paragraph 42 of *American Music* (surrounded by a circle) is linked to many paragraphs of *Folk Music*. This strong relationship turns out not to be surprising as the paragraph is a long list of "see also" references.

More interesting, however, are the links from paragraph 35 of *Folk Music* (marked with a rectangle). This paragraph discusses the changes within folk music during modern times, including primarily how it has impacted other musical genre. The paragraphs in *American Music* are brief discussions of several genre, including a mention of how folk music affected them.

Other interesting relationships between documents can be found by decomposing the documents into parts and comparing how their parts are related. Figure 4.5 is very much like Figure 4.3, except that the nodes of the graph are paragraphs of the documents rather than the entire documents. (The node annotations are omitted. Recall that when a node has no adjacent edges, it is removed from the graph: *e.g.*, the triangle in Figure 4.3f is narrower than the corresponding triangle in the previous graph because intervening nodes have been dropped.)

When a very low similarity threshold is applied, as in Figure 4.5a, the graph is far too complex to be useful to a human. Once the threshold has been raised high enough, however, several groupings of document paragraphs become apparent: in Figure 4.5f, several groups are obvious. Figure 4.6 is a larger version of the same figure with annotations. The following groups are highlighted:

- **Sousa** (extreme right of figure) is collected from a paragraph each from the *Sousa* and *American music* articles. It discusses Sousa and his effect on American music.

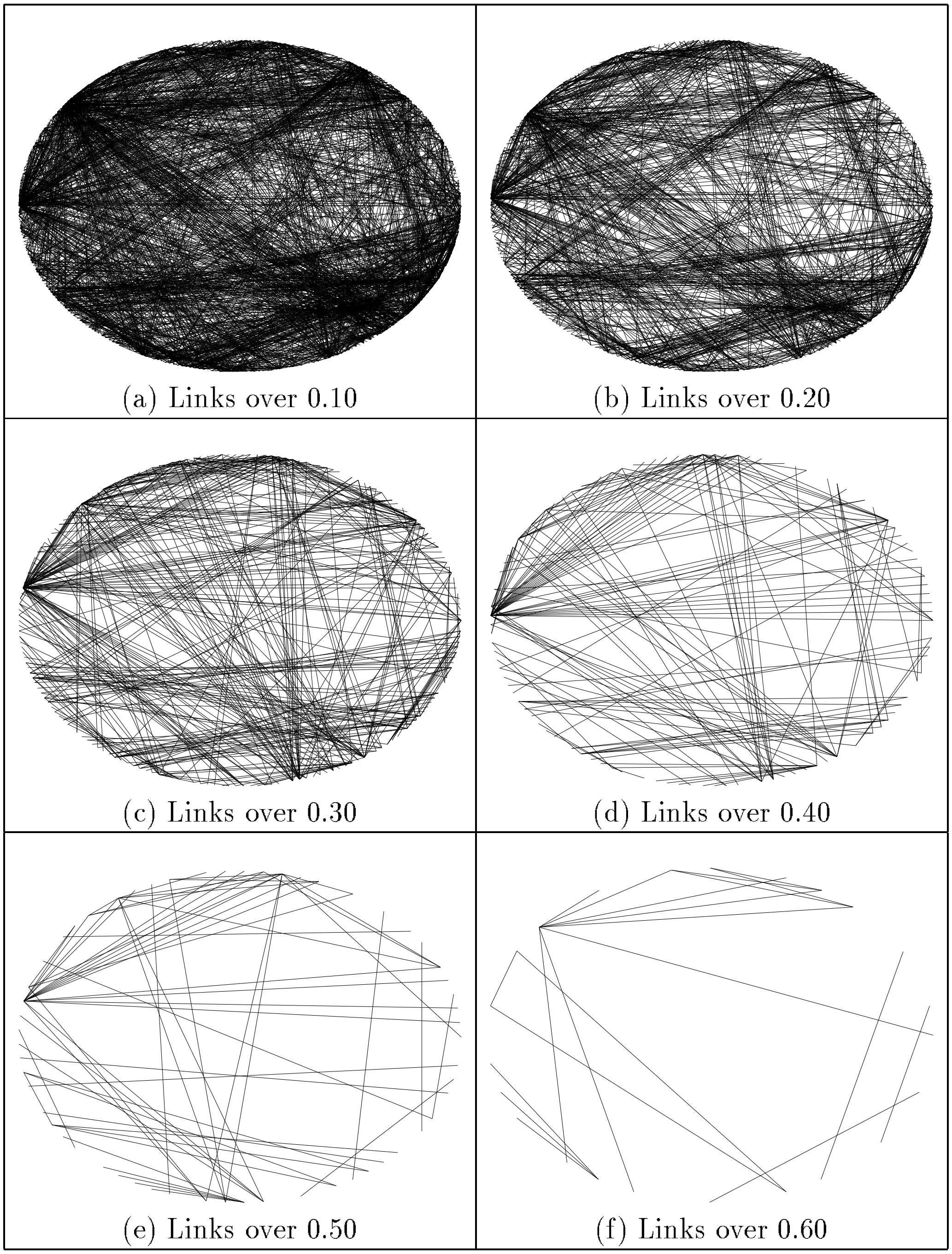
- **Indian folk music** (along the top) is an amalgam of *Folk music* and *Indian music*.

- **Sonata** is gathered from the *Western music* and *Sonata* articles.

- **Sonata in 17th century** is more specific than the general sonata discussion, and comes from the *Sonata* and *Chamber music* articles.

- **Rhythm** (bottom left) is primarily from the *Rhythm* article, but is strongly connected to the discussion of rhythm in *Popular music*.

- **"See also" pointers** is a set of links fanning out from a paragraph in the *Western music* article which refers the reader to several other documents (most of which are links in this figure).



(a) Links over 0.10

(b) Links over 0.20

(c) Links over 0.30

(d) Links over 0.40

(e) Links over 0.50

(f) Links over 0.60

Figure 4.5: Raising similarity threshold on paragraphs

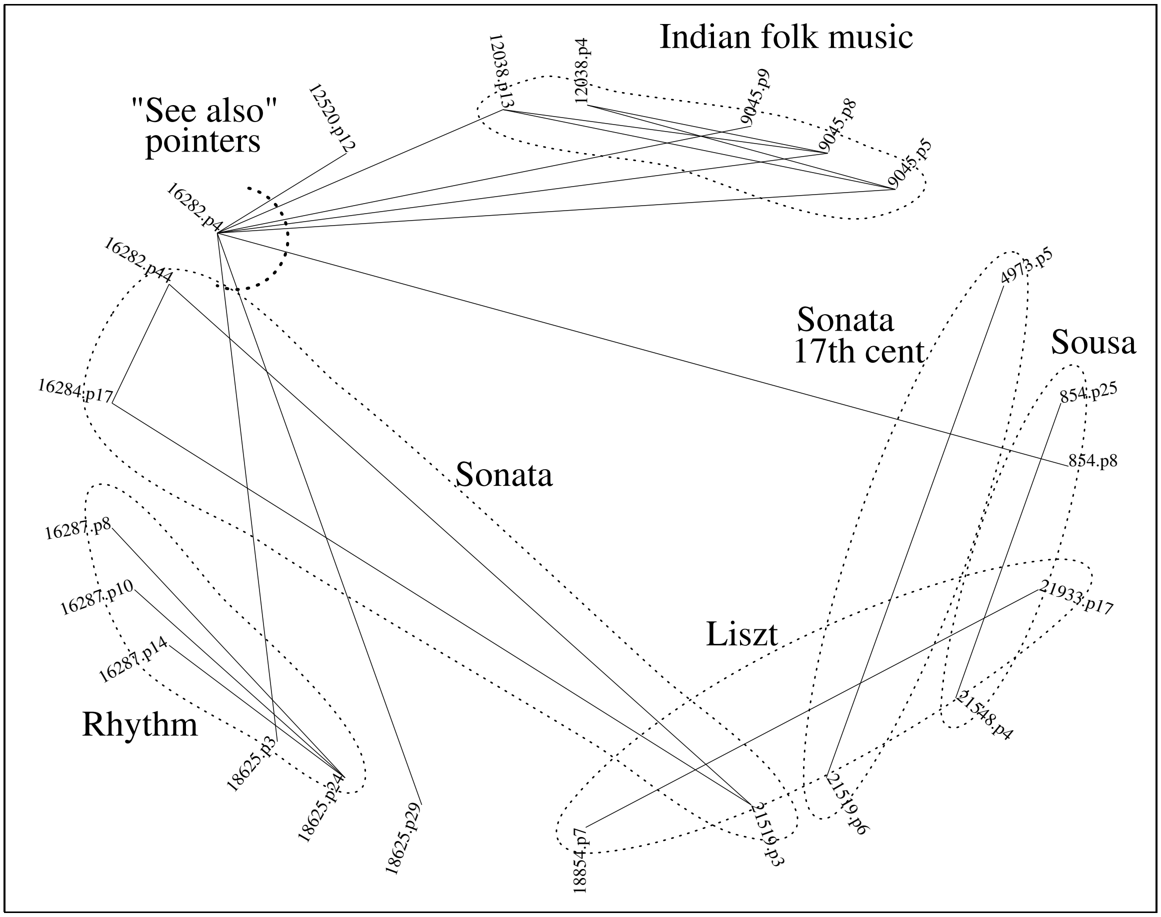


Figure 4.6: Paragraph groups within *Music* articles

Note that, in the same way that the *March music* theme disappeared when the threshold were raised for entire documents, that theme also disappeared here. (However, the *March music* theme was still present at the 0.50 threshold of paragraphs, whereas it was eliminated at that level in the document thresholding.)

This technique of slowly raising the similarity to find groups of document parts which are strongly related has been used as the basis for theme extraction and other summarizing techniques.[SABS94,SS94]

4.1.2 Uniform visual of single document

The uniform visuals provide interesting information about how groups of documents are related. It can also be used in an effort to understand how a single document is constructed. As an example, consider the documents of Figure 4.7a and 4.7b. In both cases, all paragraphs of a document were compared to all other paragraphs of the same document. Any paragraph pair which has a similarity over 0.25 is displayed with an edge between the similar paragraphs.

Although both documents are roughly the same size (about 20 thousand characters), and both have about 40 paragraphs of text, the internal relationship between paragraphs is substantially different. The paragraphs of *American music* in Figure 4.7a are almost unrelated to each other, whereas the relationships for *Folk music*'s paragraphs in Figure 4.7b are numerous and interwoven.

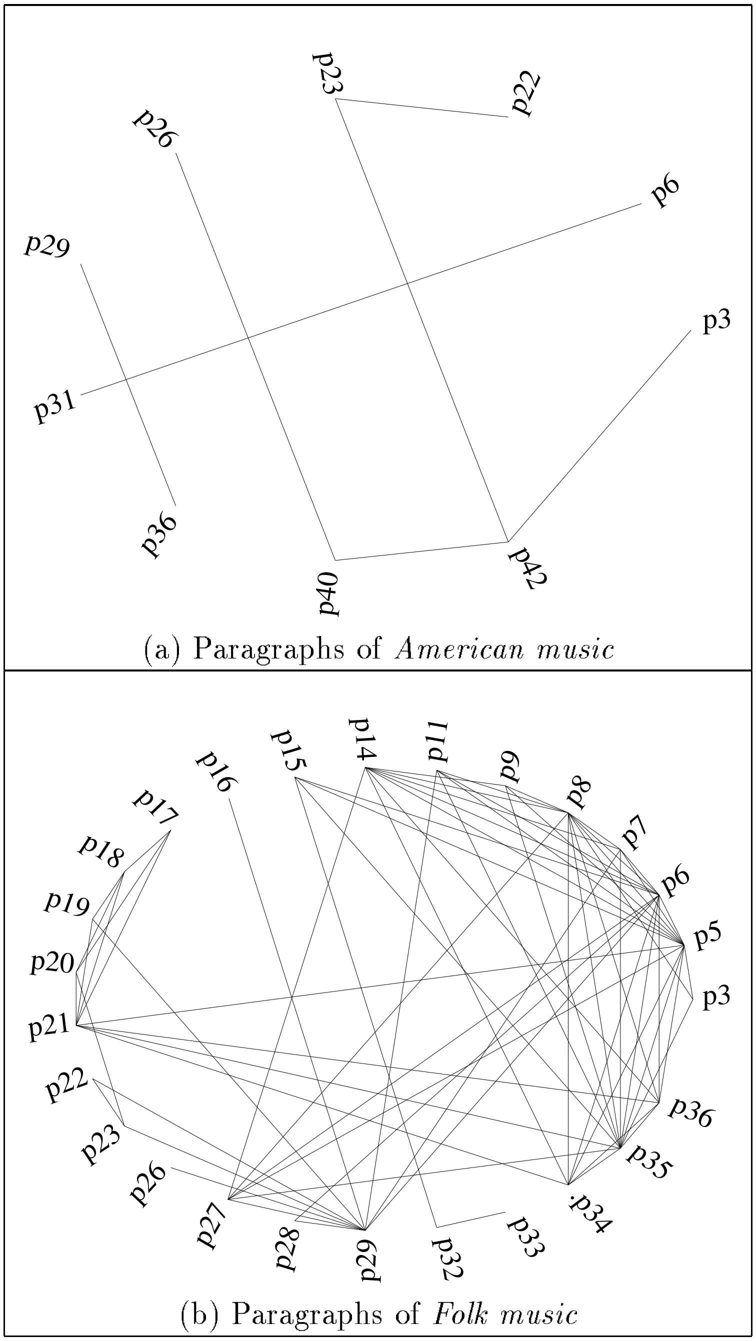
In fact, the *American music* article surveys several categories of music which—although all "American"—are otherwise only tenuously related. The *Folk music* article, on the other hand, is about a single category of music and is much more tightly focussed. The quite different styles of the articles is detectable with the uniform graph.[SABS94]

4.2 Varying with size

Figure 4.4 presented the paragraphs from both *America Music* and *Folk Music* such that each paragraph occupied the same amount of the graph and they were all equally spaced around the oval. What the graph lacks is information about how these paragraphs relate to their containing document. Figure 4.8 is an example of a different kind of visualization which provides that information in addition. The two paragraphs highlighted in Figure 4.4 are also highlighted in this new graph so that their relationship is easier to see.

This graph is composed of three elements:

1. A document is represented by a curved bar whose length is proportional to the length of the document. Figure 4.8 shows two documents, one (*American Mu-*



(a) Paragraphs of *American music*

(b) Paragraphs of *Folk music*

Figure 4.7: *American* and *Folk Music* articles

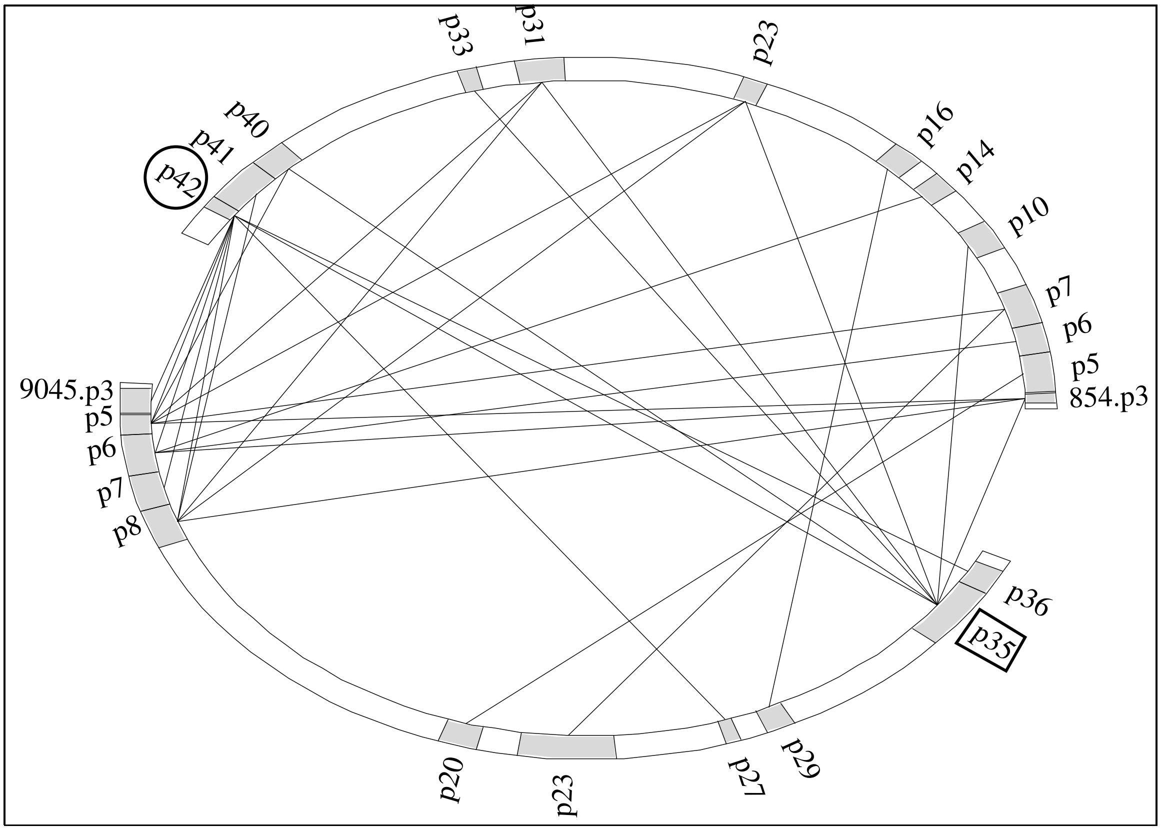


Figure 4.8: *American* and *Folk Music* articles

sic) spanning the top of the oval, the other (*Folk Music*) along the bottom. The bars are roughly the same size, indicating that the documents have roughly the same length: 21953 and 22496 characters for the top and bottom, respectively.

2. A "significant" passage of a document is shown by a shaded region in its containing document's bar. The size and position of the shaded region is proportional to the size and location of the part within its containing document.

   Documents are usually drawn assuming that the document's text spans counterclockwise around the bar. In Figure 4.8, the start of the top document is at the right and the start of the bottom document is at the left. (When only two documents are displayed, it is often preferable to have them both move left to right; most graphs will follow that convention in the rest of this work.)

3. Similarities above a chosen threshold are shown by drawing an edge between the two similar parts.

The graph is normally annotated with part numbers for the shaded regions, document numbers for the bars, and similarity measure for the edges, though these annotations are often removed manually for the sake of clarity.

The graph of Figure 4.8 was automatically generated by:

1. Breaking the documents into component parts (paragraphs in this case);

2. Finding significant paragraph pairs as discussed in Section 2.3;

3. Drawing the document bars and shading significant paragraphs (those that have a link); and,

4. Connecting the linked paragraphs.

The graphs in this thesis will frequently be hand-edited for the sake of readability. In almost all cases that means only that text will be abbreviated, enlarged, and moved.

This visualization method is new within the field of Information Retrieval, but similar techniques developed independently have been used in database analysis.[Dav93]

4.3 Visualizing relationships

This section presents several pairs or groups of documents for which a relationship is apparent from one of the styles of visuals discussed above. Each of these relationships is obvious to a human observer. Chapter 5 will examine methods which allow a computer to automatically detect the relationships as well.

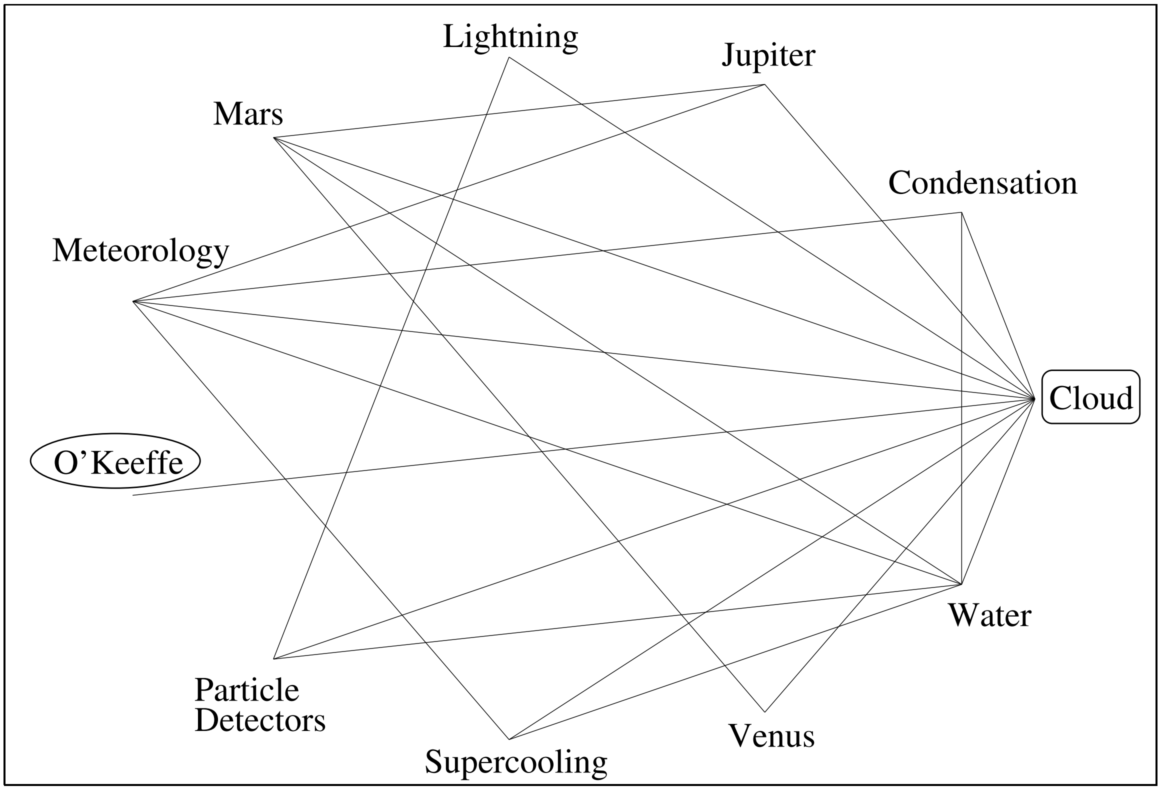


Figure 4.9: Documents related to *Clouds*

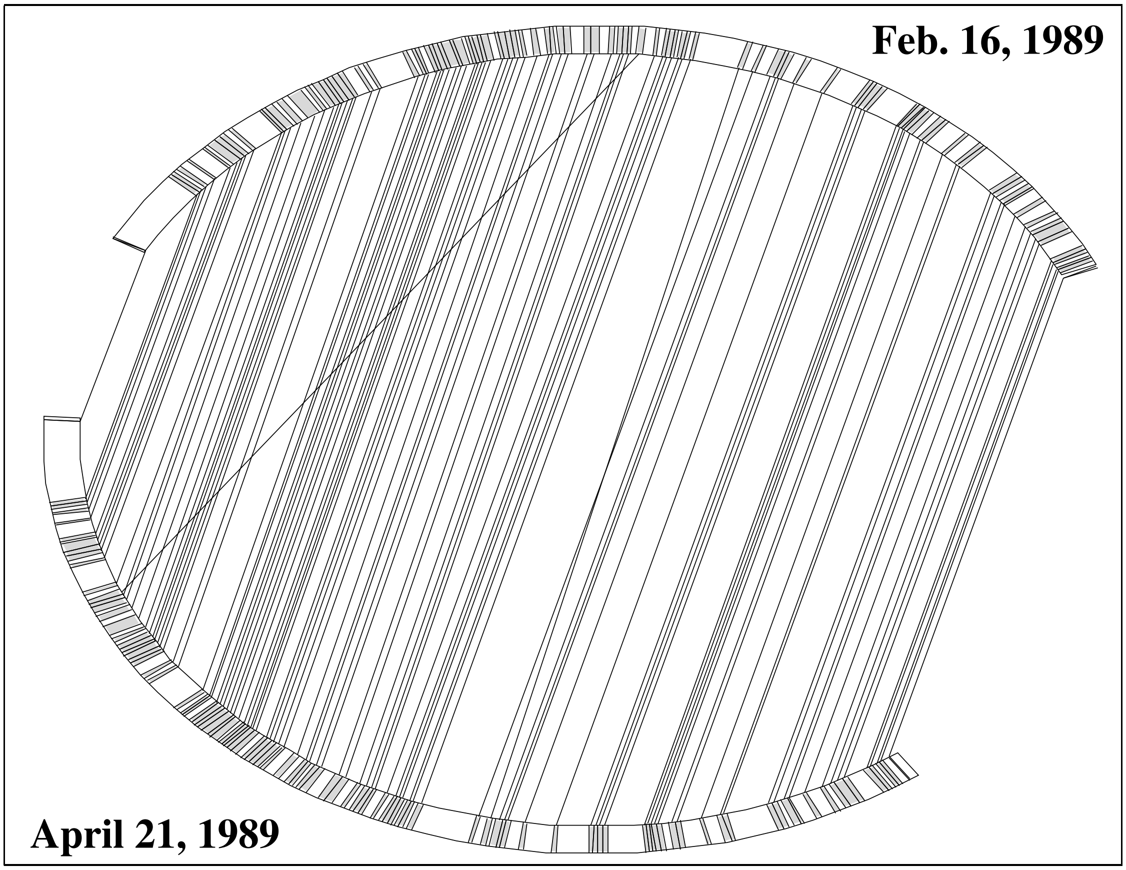


Figure 4.10: A document and a modified version

In Figure 4.9, a group of 10 documents related to the topics discussed in the article *Clouds* are presented. One document stands out as unrelated to the others: Georgia O'Keeffe. In 1965, O'Keeffe painted a mural entitled *Sky above Clouds* which causes the connection between the two. The unusual nature of the connection is highlighted by the contrasting number of links.

Figure 4.10 shows a pair of documents from the Federal Register. In February of 1989, a list of "orphan drugs" was published in the Federal Register. In April of the same year, the list was modified and published again. In the figure, identical paragraphs (those with a similarity of 1.0) from the documents are linked. The parallel nature of the links is strongly indicative of their relationship.

The same cumulative list of "orphan drugs" was published in January 1988, but an error in the data entry caused the list to be arbitrary split in two. Figure 4.11 shows the February 1989 list (as in Figure 4.10) and both halves of the January 1988

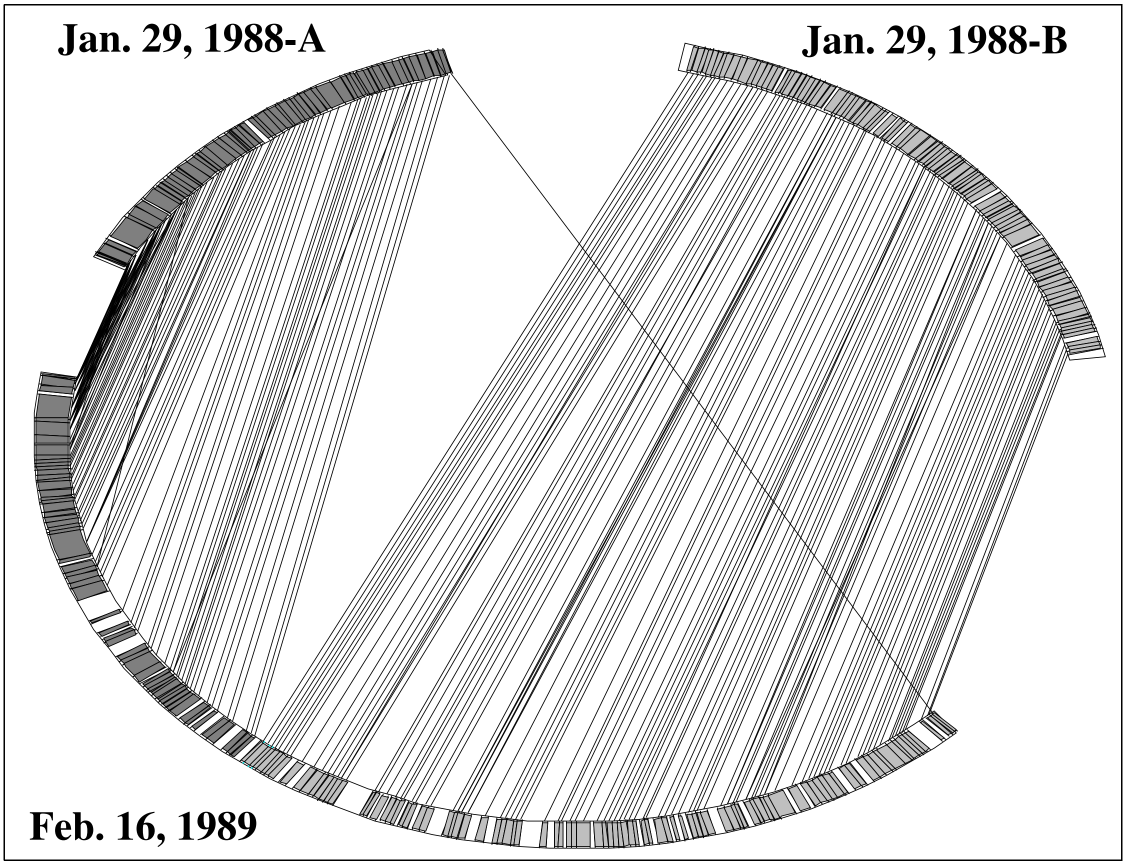


Figure 4.11: A document, modified and broken

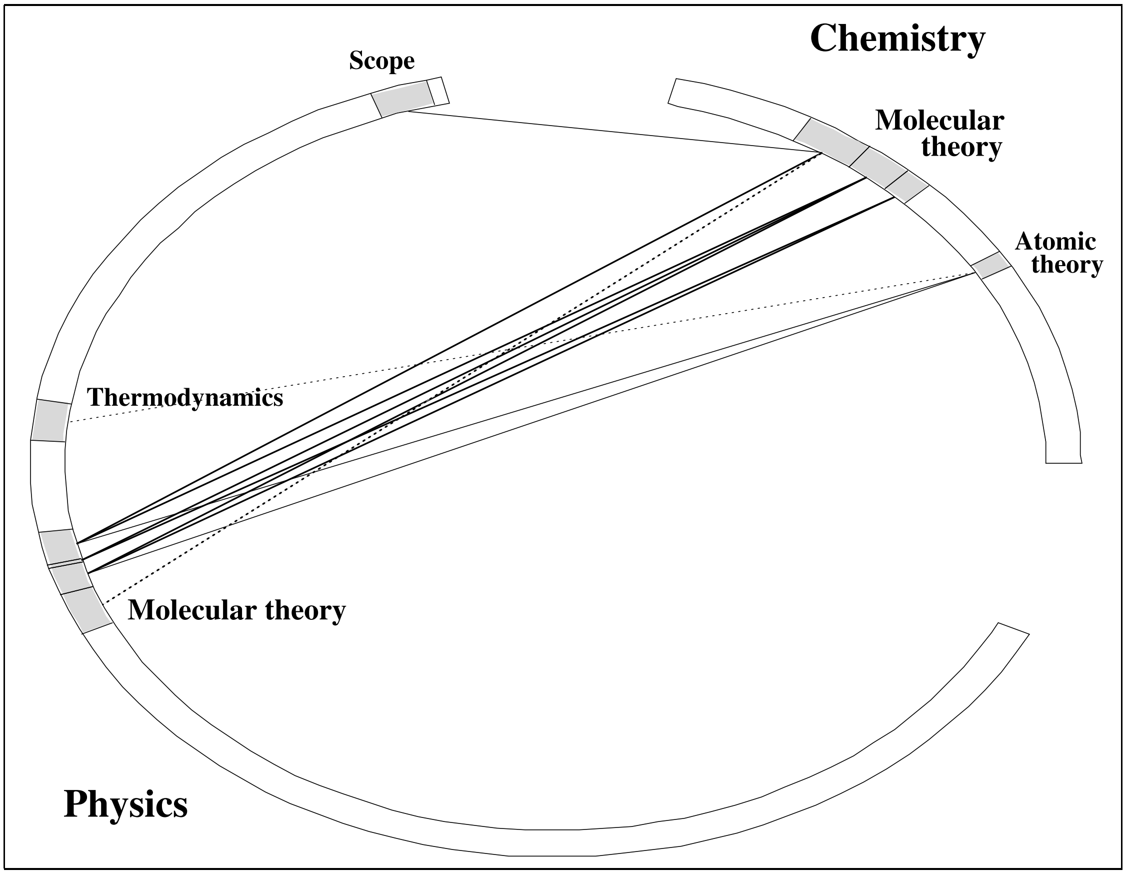


Figure 4.12: Shared subtopic of *Physics* and *Chemistry*

document. Again, only identical paragraphs are linked. The figure clearly shows the relationship between the documents.

Note that the documents in Figures 4.10 and 4.11 were all retrieved and linked automatically in response to a query. Although only selected examples are being shown, these pairs (or triples) of documents were found in the course of ordinary information retrieval.

Figure 4.12 shows how the articles on *Physics* and *Chemistry* are related. The primary relationship occurs between a series of passages roughly in the center of each document. Closer examination of the articles shows that both passages discuss molecular theory, an area where Chemistry and Physics overlap. (The highlighted sections of the document were annotated manually.)

In most cases, however, the relationship between two documents is less localized than it is in the *Chemistry* and *Physics* pair. Figure 4.13 shows the links between the

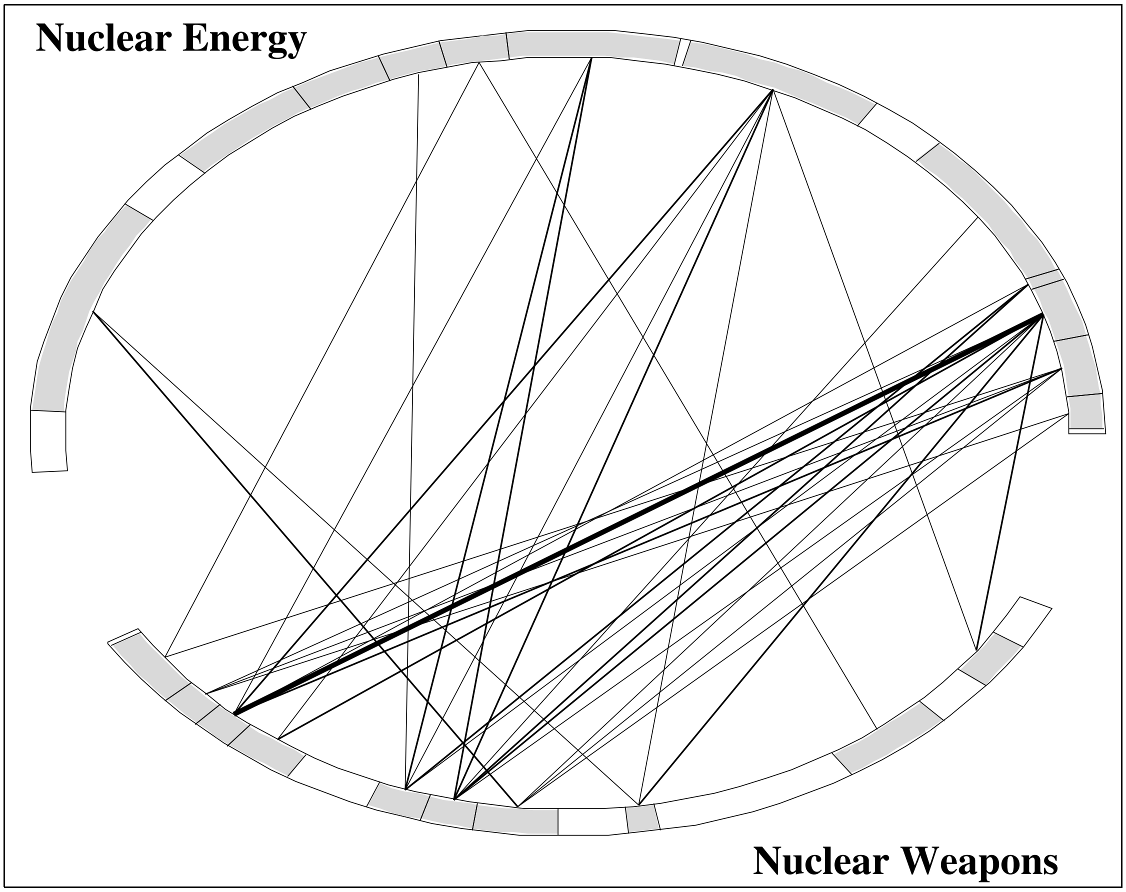


Figure 4.13: Related topics of *Nuclear energy* and *Nuclear weapons*

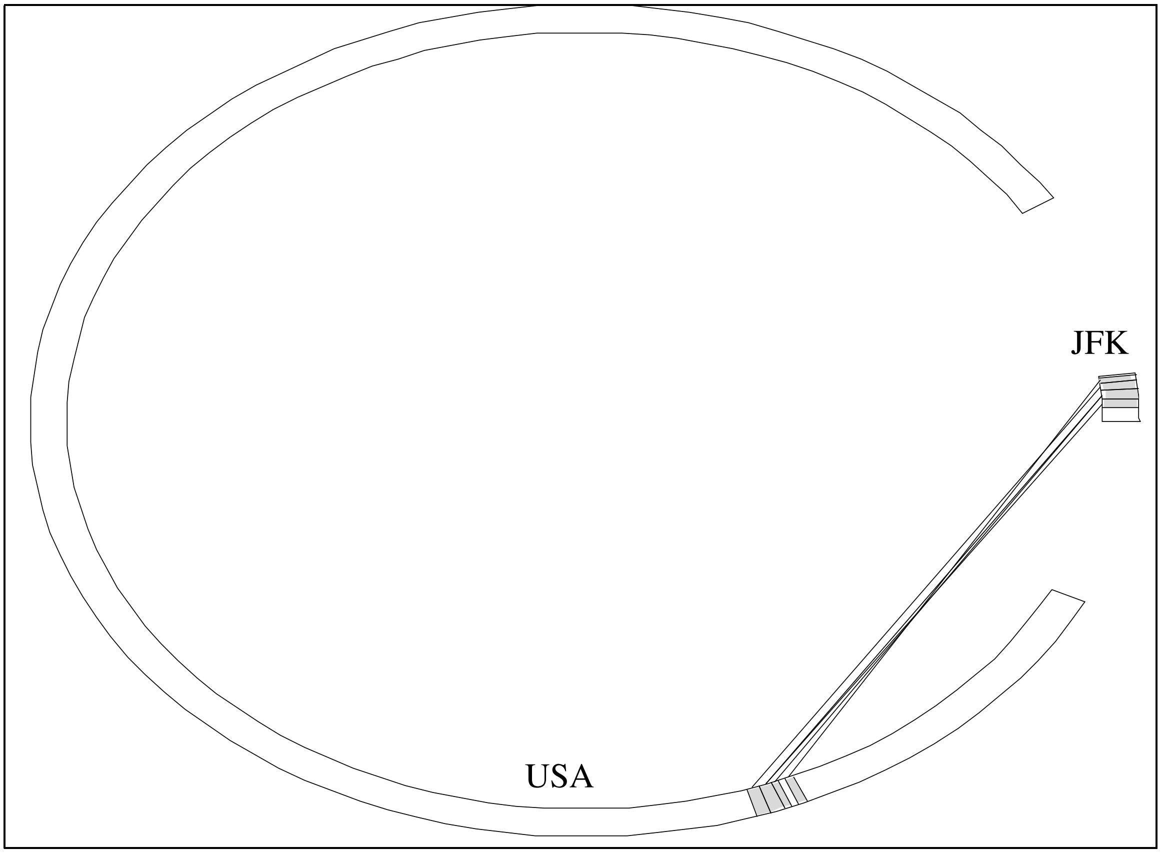


Figure 4.14: *Kennedy* as a subtopic of *United States*

documents on *Nuclear energy* and *Nuclear weapons*. In this case, almost all of each document is linked to almost all of the other. As a result, the most one can conclude is that the documents discuss similar material—a correct conclusion in this case.

In some cases, the variable visualization technique very clearly shows a relationship which would never be apparent in the uniform approach. Figure 4.14 shows the passages of an article on *John F. Kennedy* which are related to passages in an article discussing the whole of the *United States*. From this graph, it is clear that Kennedy represents a small portion of the overall discussion of the United States.

These various graphs make it clear how valuable visualization can be in recognizing the relationships between pairs or groups of documents. Inspired by these graphs, Chapter 5 will demonstrate techniques for automatically discovering some of these relationships so that they can be used to describe the type of link between the two documents.

These visualizations cannot, however, be viewed only as a means for typing links. They also provide valuable information to help a person understand how groups of documents are related. Based upon these visuals, substantial research has been done to help organize and summarize documents.[SABS94,SS94]

Chapter 5

Document Link Typing

Chapter 2 presented techniques derived from Information Retrieval for finding groups of related documents, thereby providing a means for automatically linking similar texts in a collection. In its discussion of visualization, Chapter 4 included many examples where a graphical representation of the relationship between documents or document parts suggested additional classes of relationships. As discussed in Chapter 1, it is crucial that the automatically-derived relationships also be automatically described: a large set of links between documents, without some annotations on links, will serve little purpose other than to facilitate confusion.

This chapter presents several methods for automatically identifying classes of link types. The chapter starts with a discussion of how the visualization graphs of Chapter 4 can be simplified in a systematic way to make it possible to detect link types. The chapter continues with a discussion of aggregate link types, and concludes with the taxonomy of link types (from Chapter 3) and how to identify them.

5.1 Graph simplification

The fundamentals of the procedure for automatically describing the type of the link between two documents are motivated by the global/local document similarity methods described in Chapter 2 and by the link types suggested by the visualizations of Chapter 4. The procedure is:

1. Decompose each document into smaller parts—*e.g.*, paragraphs, groups of sentences, *etc.*.

2. Compare each part of the first document to every part of the second. Remember all pairs which have non-zero similarity.

3. For each pair identified in the previous step, apply the local restriction of Section 2.3:

    (a) Break the parts into sub-parts. For example, if the documents were broken into paragraphs, this step might break the paragraphs into sentences.

    (b) Compare each of the sub-parts of the parts in a manner as above. Note the best sub-part similarity for the pair.

    If there is *at least one* sub-part pair of the part pair which exceeds a threshold (the "sub-part threshold"), identify the part pair as "good." Otherwise, if there is no such sub-part pair, mark the part pair as "tenuous."

4. Any "good" part pairs which have a similarity over another threshold (the "strong threshold") are marked as "strong"; the others are "weak."

5. Simplify the connections between the documents' parts by merging nearby part links and their incident parts.

6. Identify patterns within the simplified set of part links, and use those patterns to describe the type of the link.

As an example, consider using this approach to find the relationship between the *Physics* and *Chemistry* articles depicted in Figure 4.12 (on page 43). The first step, decomposing the documents into parts, is represented in Figure 5.1a, where the two documents are displayed with the paragraph boundaries marked. In Figure 5.1b every paragraph in the *Physics* article has been compared to every paragraph in the *Chemistry* article and all non-zero similarities have a corresponding edge drawn in the graph.

The next two portions of the figure show the results of applying steps 3 and 4. Figure 5.1c displays only "good" edges, those paragraph pairs that pass a sentence-sentence comparison test (requiring at least one sentence pair to match at a similarity of 75.0 or higher); "tenuous" links have been suppressed. Figure 5.1d then omits the "good" edges which correspond to similarities below 0.20—*i.e.*, it shows only "good" edges which are also "strong."

Finally, Figure 5.1e simplifies the graph by collapsing nearby edges and their incident endpoints (paragraphs, in this example) together. The final step is to describe the type of the main link as "contrasting treatment of subtopic."

This section on graph simplification continues by describing in more detail the document and part decomposition mentioned in steps 3a through 4, and provides additional examples as illustration. Section 5.1.2 discusses the link merging technique which simplifies the relationship between the documents substantially. Using the

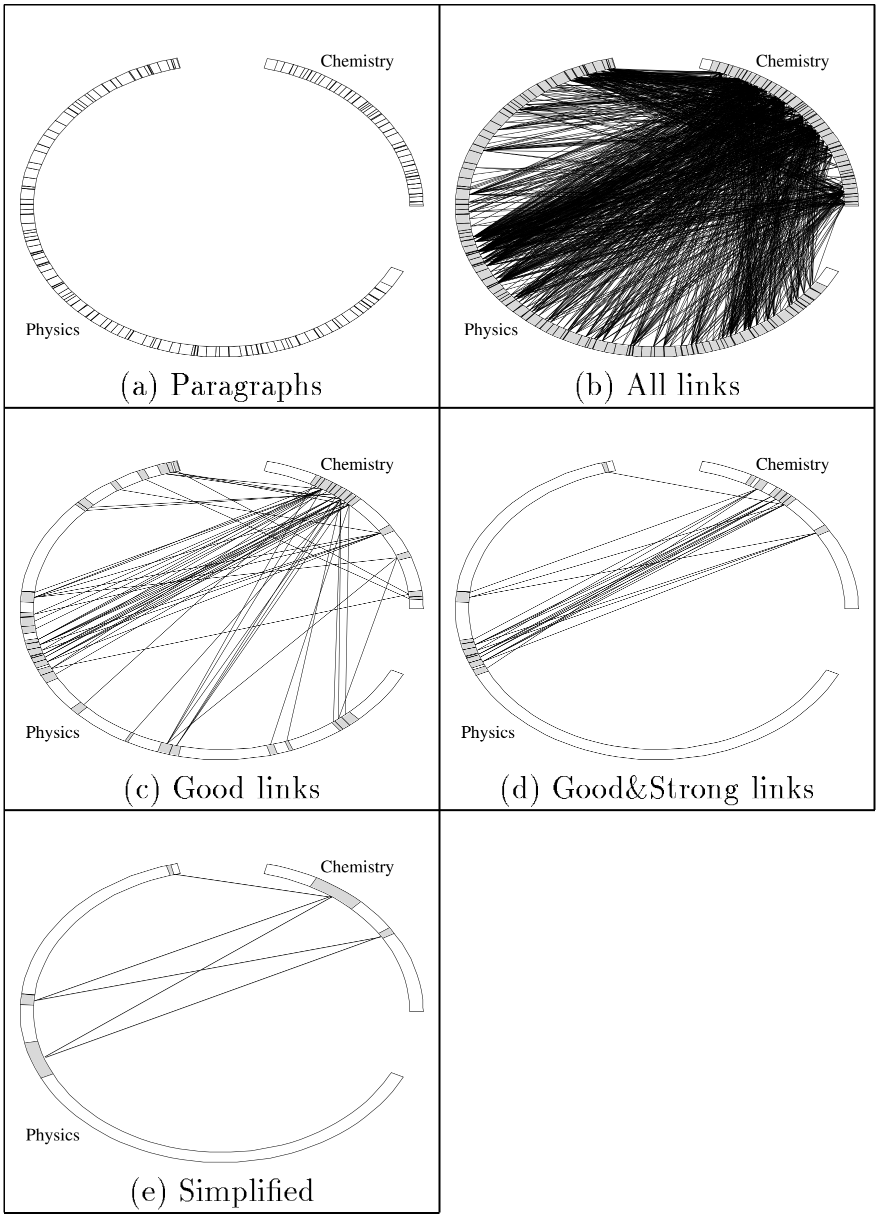


Figure 5.1: Sample typing of *Physics* vs. *Chemistry*

simplified relationship, Section 5.2 will describe the information needed to type the link and provides several examples of the technique in operation.

5.1.1 Document decomposition

Section 2.3 described a method of decomposing a pair of documents into parts and doing pairwise comparison on those parts. The intent of that comparison was to determine if the documents' vocabulary was used similarly enough to justify considering the documents relevant.

For the analysis of document-document relationship needed here, a document must be broken into parts and then the parts must be broken into sub-parts. Depending on the nature of the collection, any of the following might be possible:

Part type	Sub-part type
Section	Paragraph
Section	Sentence group
Section	Sentence
Paragraph	Sentence group
Paragraph	Sentence

(Of course, other breakdowns are possible; the listed sets serve as illustration.) For the work in this chapter, the Paragraph/Sentence breakdown was chosen as the most useful. This choice is motivated primarily because paragraph and sentence breaks can be determined automatically even in documents which are *not* annotated with markup such as SGML. (For example, paragraphs can be detected with high accuracy by scanning the text for blank or indented lines.)

The remainder of this section will assume paragraphs as parts and sentence as sub-parts, in order to make the discussion easier to follow.

After the documents have been broken into paragraphs and sentences, each paragraph of the first document is compared to every paragraph of the second document (as in Figure 5.1b). For every paragraph pair which exhibits a non-zero similarity, their sentences are pairwise compared in a similar way. The paragraph pairs are then divided into two classes:

good pairs are those which have at least one pair of sentences with a similarity above the sub-part threshold (as depicted in Figure 5.1c); and,

tenuous pairs are those which have a non-zero similarity, but do not exhibit any sentence pairs above the sub-part threshold. These paragraph pairs are suspect because they do not pass a local context check; however, they will turn out to be useful during link merging.

The "good" pairs are also divided into *strong* and *weak* pairs based upon the degree of similarity between the paragraphs—the higher the similarity, the "stronger" the pair. (Good/strong pairs are shown in Figure 5.1d.)

Given a paragraph from one document and another paragraph from the second document, then, they can have any of the following relationships:

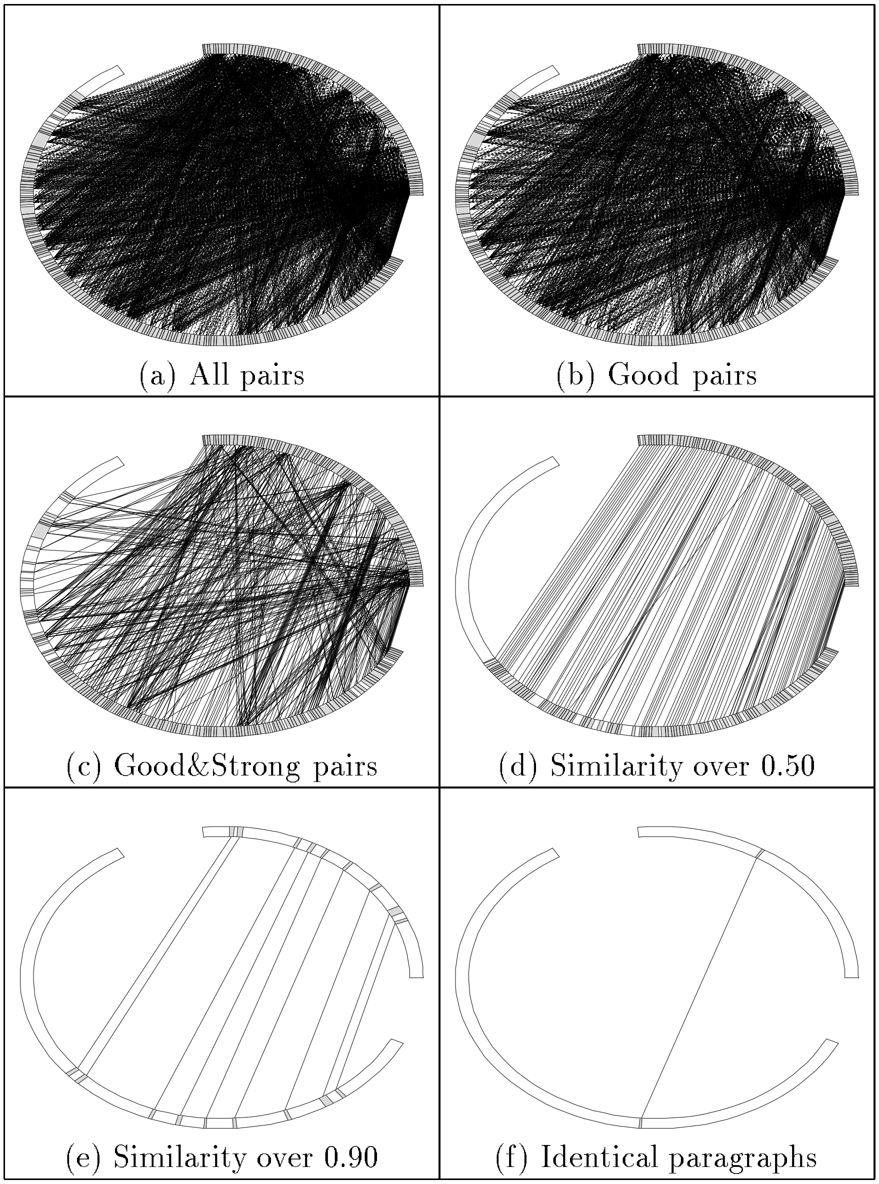
Relation	Meaning
None	Zero similarity—*i.e.*, no overlapping terms
Tenuous	No sentence pair which exceeds the sub-part threshold
Good, weak	Paragraph similarity is below the strong threshold, and there is a sentence pair which exceeds the sub-part threshold
Good, strong	Both thresholds are satisfied

5.1.2 Need for graph simplification

Figure 5.2 presents variable visualizations (see Section 4.2) of the relationship between two documents from the Federal Register in several ways (the same documents were shown in Figure 4.10). In all cases, good/strong pairs are represented by a dark line between the respective paragraphs, good/weak links by a solid line, and tenuous links by a dotted line. It is clear from the links shown in Figure 5.2a that the two documents are related, but there are far too many links to allow a more specific description of the relationship to be found.

In Figure 5.2b, the tenuous links were removed in an effort to simply the graph; unfortunately, even then there are too many links to describe the relationship. In Figure 5.2c, only the strong and good links remain, but the graph is still too dense. Figures 5.2d-f show a progressively increasing similarity threshold so that only stronger and stronger links are shown. By the time the threshold has eliminated all but the identical links (good links with similarity of 1.0) in Figure 5.2f, there are too few links to describe the relationship. It is only in Figure 5.2d that the highly parallel relationship is evident.

The process just described suggests one method of simplifying a set of part pair links in order to understand the document relationship: raise the similarity threshold until the graph is simple enough. Unfortunately, it is difficult to know exactly when "simple enough" has been achieved—after all, Figure 5.2e is simpler than Figure 5.2d, but less useful. One approach which we have used relies upon considering the "density" of the graph: the number of edges as a factor of the number of nodes. For example, low-weighted and/or tenuous edges can be dropped until there are no more than five times as many edges are there are nodes in the graph. Such a method can be


(a) All pairs

(b) Good pairs

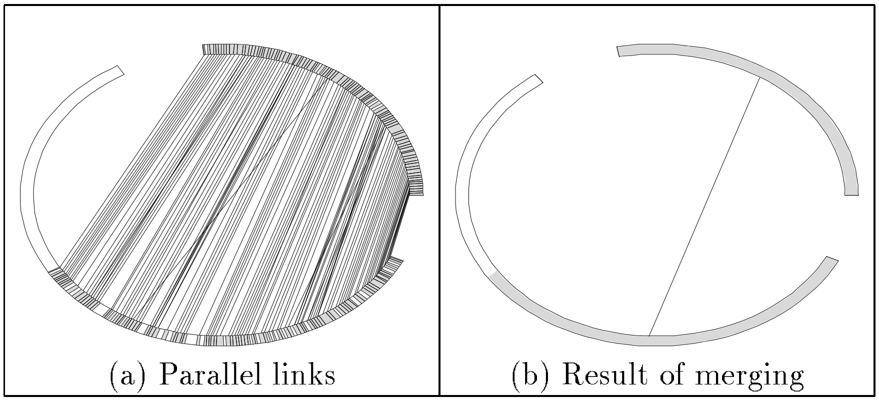
(c) Good&Strong pairs

(d) Similarity over 0.50

(e) Similarity over 0.90

(f) Identical paragraphs

Figure 5.2: Linked parts in Federal Register documents



(a) Parallel links

(b) Result of merging

Figure 5.3: Simplifying graph by merging links

quite effective, but dropping low-similarity links from the graph discards information which could be useful. It is preferable, therefore, to simplify the graph differently.

The approach we will take is motived by the graph of Figure 5.3a (which is identical to Figure 5.2d). That graph consists of many links, almost all of which are roughly parallel. The endpoints of the links span large sections of each document and are almost entirely adjacent. An obvious way to simplify that graph would be to merge the links and their endpoints to create a single "meta-link" which includes the information of all of those links. Figure 5.3b shows the result of creating such a meta-link.

5.1.3 Link merging

The process of merging links is dominated by two questions: when should links be merged, and how should the merge occur? In Figure 5.3a, the mostly parallel links and very close positioning of the endpoints clearly suggest that all of the links could be treated as one larger link. But links may be related in other ways, or they may not be as closely spaced. The choice of whether or not to merge links is based upon those two factors.

Figure 5.4 shows the four types of relationships that a pair of links might have: a *parallel* pair is one such as those depicted in Figure 5.4a (the links in Figure 5.3b are also almost entirely parallel);[1] a *wedge* pair as in Figure 5.4b requires that the links share a common endpoint; a *crossing* pair requires that the edges cross at some

[1] "Parallel" becomes difficult to define when the documents are of different sizes. In this work, the proportional lengths of documents is considered, and the center of the link's endpoint is used to decide where the lines lie. If the distances between the center-points at each endpoint are within 5% of identical, the links are considered parallel.

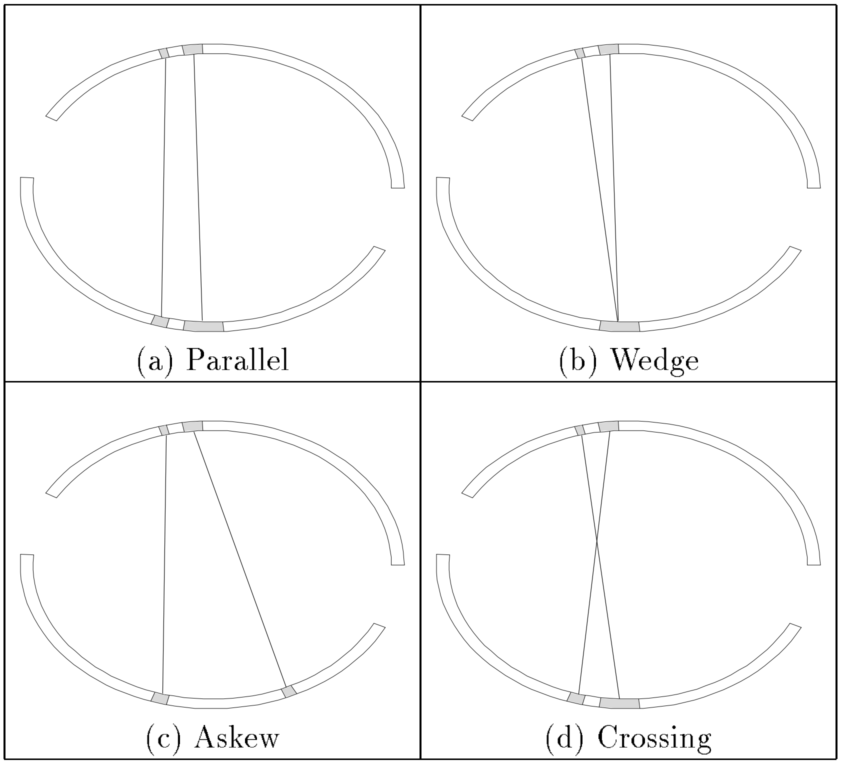


Figure 5.4: Link edge relationships

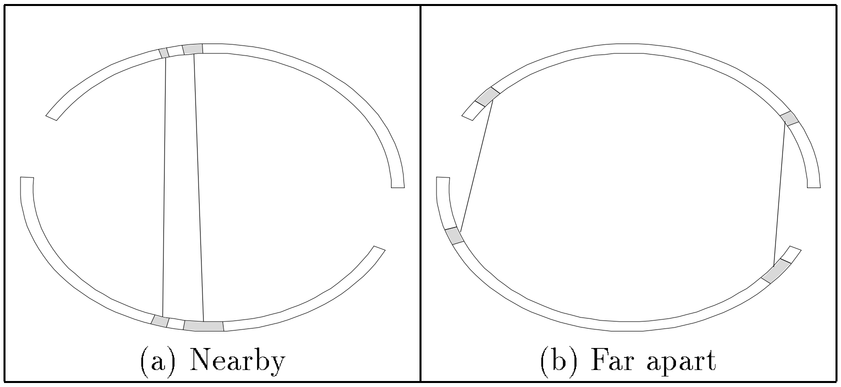


Figure 5.5: Varying distance between links

point; and an *askew* pair is one which does not cross, but which is not parallel.

Any type of pair relationship could be useful for merging, but the distance between the nodes is of obvious importance. In Figure 5.5, both of the pictured pairs are parallel (or nearly so), but though it is reasonable to consider merging the pair in Figure 5.5a, the pair of 5.5b would be merged only as a last resort.

Figure 5.6 shows an abstract representation of two links, A and B, between parts of documents which have sizes σ_1 and σ_2. The nodes incident on link A have sizes α_1 and α_2, and the nodes incident on link B have sizes β_1 and β_2. δ_1 and δ_2 are the sizes of the *unlinked* text which lies between the endpoints on each end of the pair. Whether or not pair A and B should be merged depends upon the size of δ_i compared to the sizes of α_i and β_i.

In this work, the distance between the pair of links depicted in Figure 5.6 is calculated as:

$$\text{Distance} = \sum_{i=1}^{2} \frac{\delta_i}{\sigma_i} \tag{5.1}$$

That is, the distance between links is related to the proportion of the total documents which lies between their endpoints. The link pair with the smallest distance will be merged first; ties are broken by considering the relationship between the links. (A pair of links in a wedge relationship—as in Figure 5.4b—is always considered closer together than the similar askew pair with adjacent endpoints.) After a link pair has been merged, the new link will have a different distance from the remaining links. The process continues until no links are "close enough" for merging.

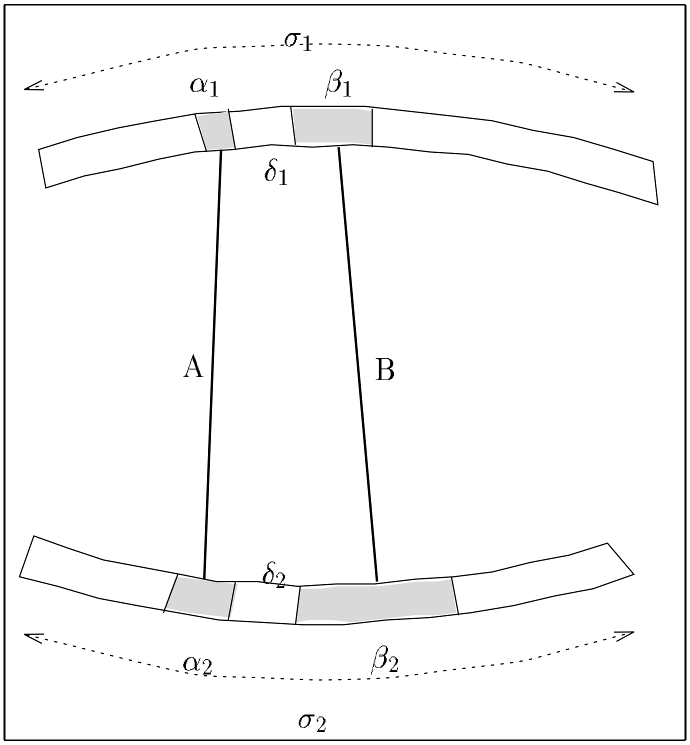


Figure 5.6: Sizes of components of a link pair

5.1.4 Similarity of meta-link

When two links are merged, the resulting meta-link is between different passages of text and may even include some previously unlinked text. The similarity assigned to that edge must therefore change.

It would be possible to recompute the similarity by re-calculating the vectors for the merged endpoints and computing the similarity directly, but it is possible to achieve a similar effect by combining the similarities appropriately. The new similarity must satisfy some simple properties in order to be useful. For example, if previously-unlinked text is merged into the meta-link, the new similarity will necessarily be lower than the original. Appendix A describes properties expected of a new similarity, presents the formula used to calculate the new value, and proves that the formula satisfies the properties.

5.1.5 Stopping the merge

The link merging process requires some means for knowing when merging should stop. Figure 5.5b pictured a situation where the link merging almost definitely should not happen because the links are too far apart, but there are other possible stopping criteria:

- Merge *all* of the links between the parts of two documents;
- Merge until no remaining links are above a threshold (every merged "meta-link" has a similarity lower than the links used to create it); or,
- Merge until the "distance" between the closest link pair is larger than some threshold.

The first option, merging all of the links so there is only a single meta-link, seems useful in that it gives an indication of the overall similarity between the two documents. However, the value is somewhat dubious since an overall document-document similarity can be calculated much more efficiently by considering the document vectors (as in Section 2.1.2).

The second choice, stopping when no links have a strong similarity (or only one link exists), will in many cases result in merging links that are too separated to be very related. For example, the "far apart" links of Figure 5.5b might be merged if this approach is used.

The final choice is to merge until the distance between the closest link pair is larger than some threshold. Intuitively, this choice means that the "nearby" links of Figure 5.5a would be merged, but the "far away" links of 5.5b would not. This criterion is the one used in this work.

The "distance" between a pair of links is defined by Equation 5.1, as the sum of the proportions of unlinked document which would be merged into the link. This method is appropriate because it stops the merging when the "dilution" effect of merging links becomes too great. The actual threshold can vary from 0 to 2, depending on the needs of the application. In this work, the threshold was chosen as 0.10, or roughly 5% of each document in the average case. For example, if the endpoints of a pair of links are separated by 4% of one document's length and by more than 6% of the other's, then the link pair is *not* eligible to be merged.

In the *Physics* and *Chemistry* example of Figures 5.1d and 5.1e (page 49), this threshold is what allows the tightly-clustered set of links to be merged, but prevents the three outlying paragraphs from being merged into that cluster of linked text.

5.2 Link patterns

After the links between the documents have been simplified, they can be analyzed to detect patterns which are useful for describing the link. The following values for each meta-link are useful for identifying patterns.

Convolution describes how "parallel" the links that make up a meta-link were;

Expansion describes how extraneous (unrelated) text was added in one endpoint but not another;

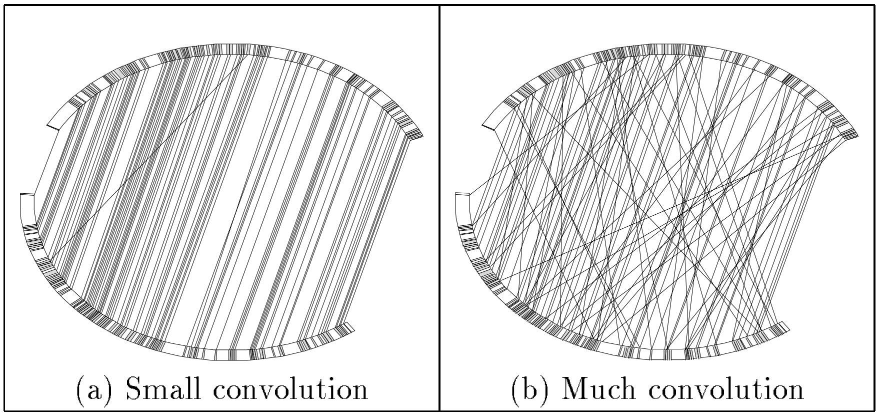
Relative size highlights the importance of the link within each of the documents; and,

Absolute size reflects the importance of the link's endpoints to each other.

The next sections describe each of those points in more detail.

5.2.1 Convolution

If a meta-link is created between two regions of text by combining several links between parts within the regions, we say the the regions paraphrase each other. We can get an even better sense of how well they paraphrase one another by a closer look at the parts which were merged. Consider the examples of Figure 5.7 which portray two regions which might be merged by a meta-link. In the case of Figure 5.7a (a genuine pair of documents from the Federal Register), the similarities follow a clear pattern, suggesting that the two regions of text follow roughly the same structure. On the other hand, Figure 5.7b (a hypothetical pair of documents) portrays two passages that have dramatically different styles of exposition.


(a) Small convolution

(b) Much convolution

Figure 5.7: Convolution of links

It is possible to measure the amount of convolution that exists in a meta-link by counting the number of crossing links which are merged. If only parallel links are merged, the meta-linked regions will have parallel structure. If only crossing links are merged, the regions will have quite varying structure.

5.2.2 Expand/condense

To keep the meta-link structure simple, when two links are merged and there is extraneous text between the endpoints of the link, that text is folded into the meta-link. Every time that is done, the amount of unrelated text included in the link increases.

Consider the meta-link of Figure 5.8 (a conference paper and a journal version of the same paper). The shaded portions of text are the the ones which were at the endpoint of a link; the unshaded portions represent extraneous text which was added when the meta-link was created. Since the unshaded portion of the top document is quite large, the amount of extraneous text—not related to anything in the bottom document—clearly detracts from the overall relatedness of the two regions.

In a case such as this, we claim that the discussion in the bottom document is a condensed version of the discussion in the top document. Conversely, that the document's treatment expands upon that of the bottom document. (Note that it is not actually reasonable to claim that the regions expand or condense one another unless they are *also* sufficiently similar to warrant considering them paraphrases of one another.)

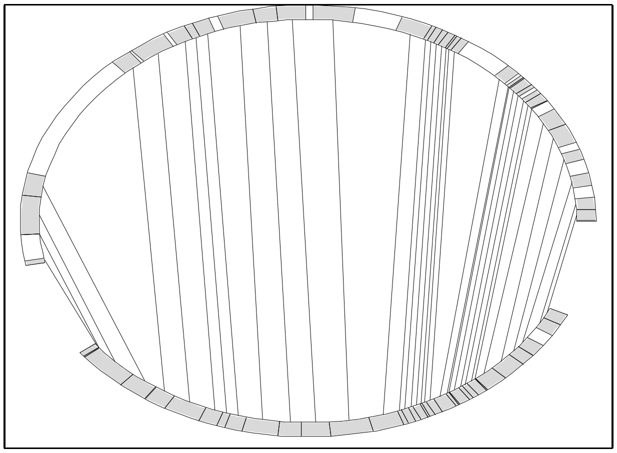
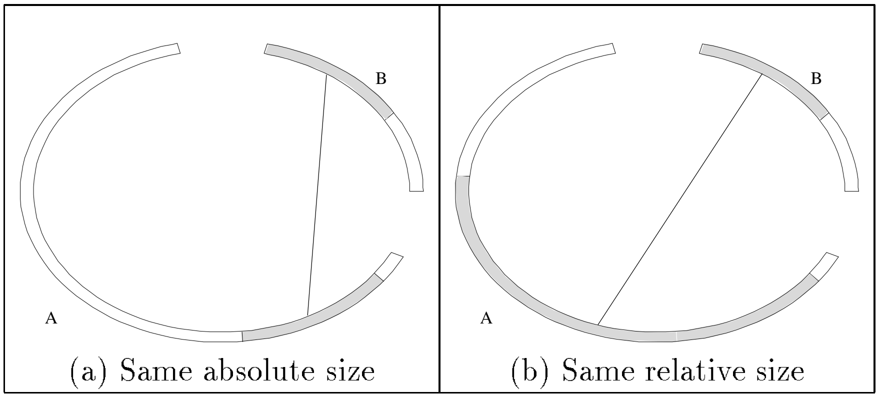


Figure 5.8: Extraneous text in meta-link



(a) Same absolute size

(b) Same relative size

Figure 5.9: Size of link endpoints

5.2.3 Relative and absolute sizes

Another important measure of a meta-link is how significant the linked text is in the respective documents. Figure 5.9a shows a hypothetical example of a link where the ends are the same size, but the linked text clearly has more significance in text B than in text A. On the other hand, Figure 5.9b shows a link where the endpoints take up the same amount of space in their respective documents (half), but are clearly *not* the same absolute size.

When considering the importance of a meta-link, both the comparative sizes of the two regions—is the amount of material covered roughly the same in both cases—and their sizes relative to their containing documents—how significant is the region within its document—are considered.

5.2.4 Describing link patterns

The aspects described above of the relationship between two documents are sufficient to construct a reasonably good picture of the type of link. In order to provide a more useful description of the link than "paraphrase," we automatically generated more detailed descriptions. The following descriptions were each automatically generated by Smart (they have been re-formatted slightly for presentation).

- The text in the final fifth of 12157 *Intelligence* (p5) is also discussed in the similar passage of increased significance (but greater coverage) in the latter half of 22285 *Lewis Madison Terman* (p4).

  Figure 5.10 shows the selected paragraphs and a variable visual of the two documents with the paragraphs linked.

- The text in the starting half of 4868 *Central Intelligence Agency, History and Activities* (p3,p5-6,p8-9) is discussed differently in the similar passage of less significance (and less coverage) in the final fifth of 6951 *Department of Defense, Supporting Agencies* (p11).

  Figures 5.11 and 5.12 show the selected text of the CIA and Defense documents, respectively.

- The text in all of 2205 *Cumulative List of Orphan Drug and Biological Designations, Jan 29, 1988* (p8-14,p18-72) is covered similarly in the almost identical passage of less significance in starting half of 29806 *Cumulative List of Orphan-Drug and Biological Designations, April 21, 1989*(p8-14,p18-64,p66-94).

  (These documents are two of the three depicted in Figure 4.11 (page 42) where a Federal Register article was arbitrarily split in the middle.) Figures 5.13 and

Text from 12157, *Intelligence*, paragraph 5:

> In the formulation of intelligence tests, most psychologists tend to adopt an eclectic concept, according to which intelligence is treated as a general ability operating as a common factor in a wide variety of special aptitudes. It is observed and measured by techniques focused upon these aptitudes singly or in combination.

Text from 22285, *Lewis Madison Terman*, paragraphs 4:

> Terman became prominent for his specialized research in intelligence testing and educational experiments with intellectually gifted children (see Intelligence). He devised the term intelligence quotient (IQ), which became an index of measurement of the intelligence level of both children and adults, with a normal standard of 100. He also developed the so-called Stanford-Binet intelligence tests to measure the IQ. Among his many important works are The Measurement of Intelligence (1916), The Intelligence of School Children (1919), The Stanford Achievement Test (1923), The Gifted Child Grows Up (1947), and The Gifted Group at Mid-Life (1959).

The two documents with passages highlighted:

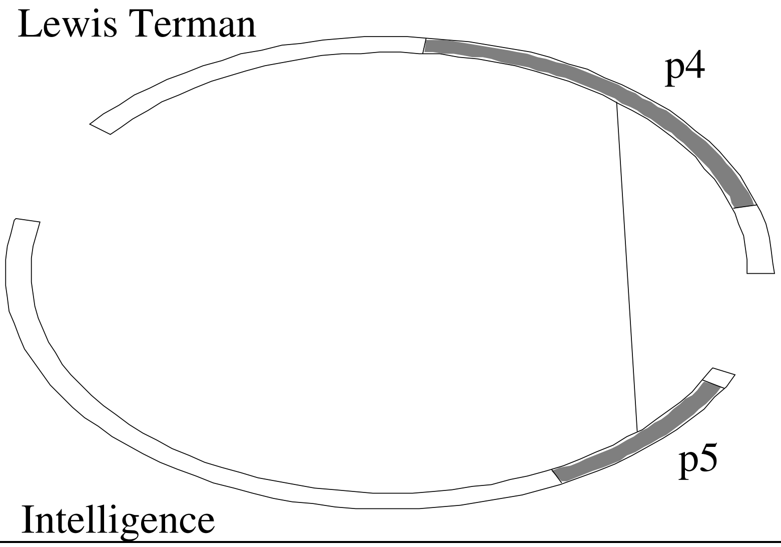
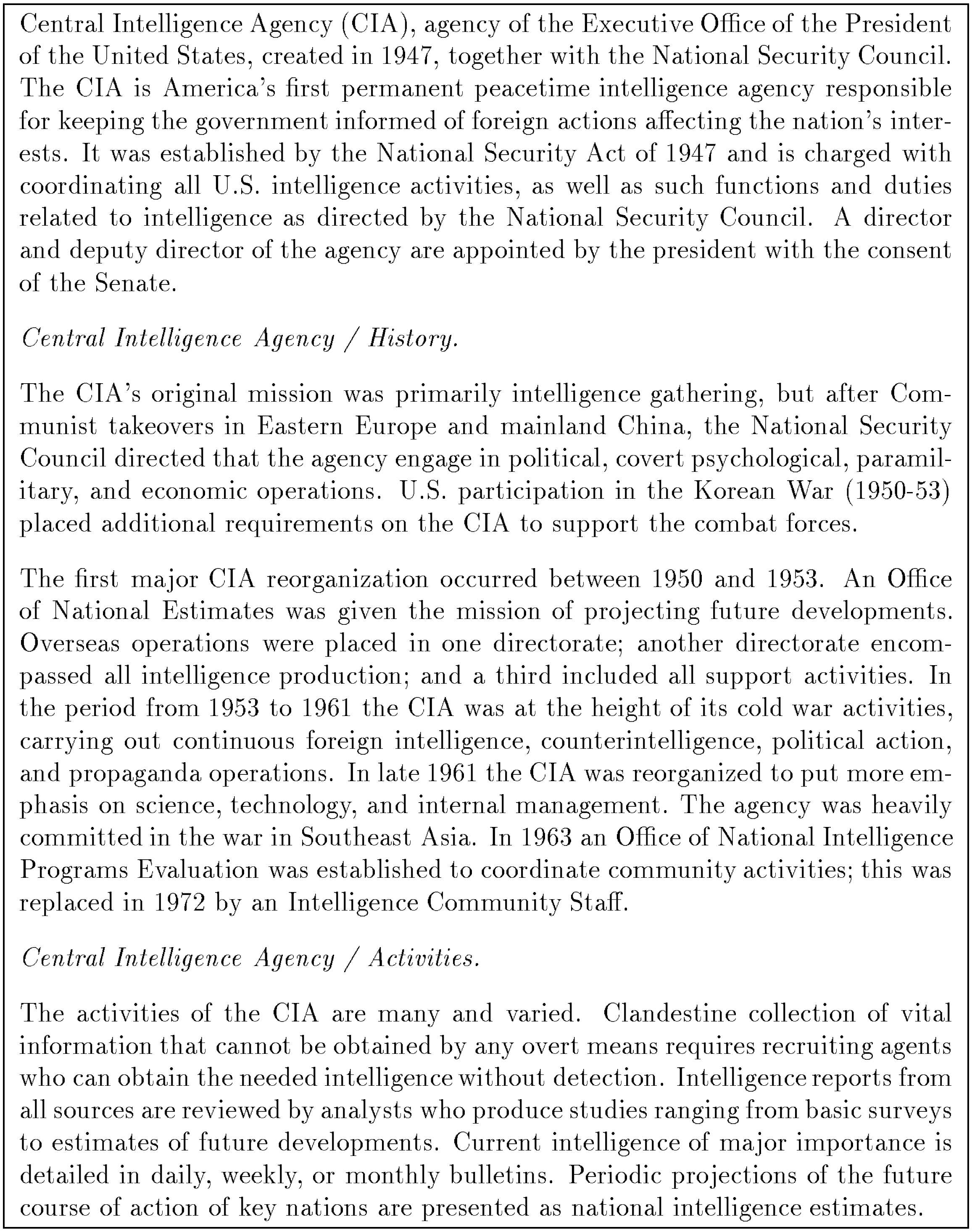


Figure 5.10: *Intelligence* and *Lewis Terman*



Central Intelligence Agency (CIA), agency of the Executive Office of the President of the United States, created in 1947, together with the National Security Council. The CIA is America's first permanent peacetime intelligence agency responsible for keeping the government informed of foreign actions affecting the nation's interests. It was established by the National Security Act of 1947 and is charged with coordinating all U.S. intelligence activities, as well as such functions and duties related to intelligence as directed by the National Security Council. A director and deputy director of the agency are appointed by the president with the consent of the Senate.

Central Intelligence Agency / History.

The CIA's original mission was primarily intelligence gathering, but after Communist takeovers in Eastern Europe and mainland China, the National Security Council directed that the agency engage in political, covert psychological, paramilitary, and economic operations. U.S. participation in the Korean War (1950-53) placed additional requirements on the CIA to support the combat forces.

The first major CIA reorganization occurred between 1950 and 1953. An Office of National Estimates was given the mission of projecting future developments. Overseas operations were placed in one directorate; another directorate encompassed all intelligence production; and a third included all support activities. In the period from 1953 to 1961 the CIA was at the height of its cold war activities, carrying out continuous foreign intelligence, counterintelligence, political action, and propaganda operations. In late 1961 the CIA was reorganized to put more emphasis on science, technology, and internal management. The agency was heavily committed in the war in Southeast Asia. In 1963 an Office of National Intelligence Programs Evaluation was established to coordinate community activities; this was replaced in 1972 by an Intelligence Community Staff.

Central Intelligence Agency / Activities.

The activities of the CIA are many and varied. Clandestine collection of vital information that cannot be obtained by any overt means requires recruiting agents who can obtain the needed intelligence without detection. Intelligence reports from all sources are reviewed by analysts who produce studies ranging from basic surveys to estimates of future developments. Current intelligence of major importance is detailed in daily, weekly, or monthly bulletins. Periodic projections of the future course of action of key nations are presented as national intelligence estimates.

Figure 5.11: Text of *Central Intelligence Agency*
(paragraphs 3 through 9)

Text from 6951, *Department of Defense*, paragraph 11:

> A number of agencies assist the Defense Department. The National Security Agency/Central Security Service performs specialized technical and coordinating functions relating to national security. The Defense Nuclear Agency provides support to all the armed forces in matters related to nuclear weapons, their effects and testing. The Defense Communications Agency is responsible for operational and management direction of a worldwide communications system for the armed forces and other government agencies. The Defense Intelligence Agency organizes and directs Defense Department intelligence resources and reviews and coordinates intelligence functions assigned to the military departments. The Defense Logistics Agency provides supply and service support for items common to all services; this includes procurement of materiel, operation of a wholesale distribution system, and surplus-disposal programs. Other supporting agencies are the Defense Contract Audit Agency, Defense Advanced Research Projects Agency, Defense Investigative Service, Defense Mapping Agency, Defense Legal Services Agency, Defense Security Assistance Agency, Strategic Defense Initiative Organization, and the On-Site Inspection Agency.

The two documents with passages highlighted:

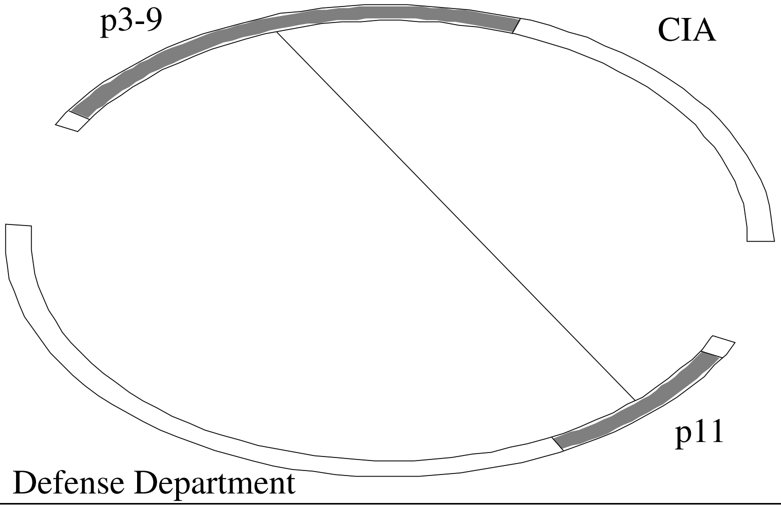


Figure 5.12: *Defense Department* compared to *CIA*

Orphan Designations Pursuant to Section 526 of the Federal Food, Drug, and Cosmetic Act as Amended by the Orphan Drug Act (Pub. L. 97-414) Through 1987 3,L2,p8,8/9,i1,s50,r100,r50 Orphan Drug and Biological Designations Through 1988 [Approved for Marketing*] [Exclusive Approval**] Biological Designations Name of biological Designated use Sponsor's name and address Generic-alpha-1-anti-trypsin (recombinant DNA origin). Trade_Not established Supplementation therapy for alpha-1 antitrypsin deficiency in the ZZ phenotype population Cooper Biomedical, Inc., 3145 Porter Drive, Palo Alto, CA 94304.

Generic-alpha-1-proteinase inhibitor (Alpha-1 PI). Trade_Prolastin */** Replacement therapy in the Alpha 1 PI congenital deficiency state Cutter Laboratories, P.O. Box 1986, Berkeley, CA 94701.

Generic-anti-J5mAb. Trade_Not established Treatment of patients with gram-negative bacteremia which has progressed to endotoxin shock Centocor, Inc., 244 Great Valley Parkway, Malvern, PA 19355.

Generic-antimelanoma antibody XMMME-001-RTA Trade_Same as generic Treatment of Stage III melanoma not amenable to surgical resection Xoma Corporation, 3516 Sacramento Street, San Francisco, CA 94118.

Generic-Anti-TAP-72 immunotoxin. Trade_XOMAZYME-791 Treatment of metastatic colorectal cancer adenocarcinoma XOMA Corporation, 2910 Seventh Street, Berkeley, CA 94701.

Figure 5.13: Text from 2205, *List of Orphan Drug and Biological Designations, January 29, 1988*, paragraphs 18 through 22

3,L2(,,0),p8,8/9,i1,s75,r75,r75 Orphan Drug and Biological Designations Through 1988 [Approved for Marketings*] [Exclusive Approval**] Biological Designations Name of biological Designated use Sponsor's name and address Generic_alpha-1-anti-trypsin (recombinant DNA origin) Trade_Not established Supplementation therapy for alpha-1 antitrypsin deficiency in the ZZ phenotype population Cooper Biomedical, 3145 Porter Drive, Palo Alto, CA 94304.

Generic_alpha-1-proteinase inhibitor (Alpha-1 PI) Trade_Prolastin */** Replacement therapy in the Alpha 1 PI congenital deficiency state Cutter Laboratories, P.O. Box 1986, Berkeley, CA 94701

Generic_anti-J5mAb Trade_Not established Treatment of patients with gram-negative bacteremia which has progressed to endotoxin shock Centocor, Inc., 244 Great Valley, Malvern, PA 19355.

Generic_antimelanoma antibody XMMME-01-RTA Trade_Same as generic Treatment of Stage III melanoma not amenable to surgical resection XOMA Corporation, 3516 Sacramento St., San Francisco, CA 94118.

Generic_anti-TAP-72 immunotoxin Trade_XOMAZYME-791 Treatment of metastatic colorectal adenocarcinoma XOMA Corporation, 2910 Seventh Street, Berkeley, CA 94710.

Figure 5.14: Text from 2205, *List of Orphan Drug and Biological Designations, April 21, 1989*, paragraphs 18 through 22

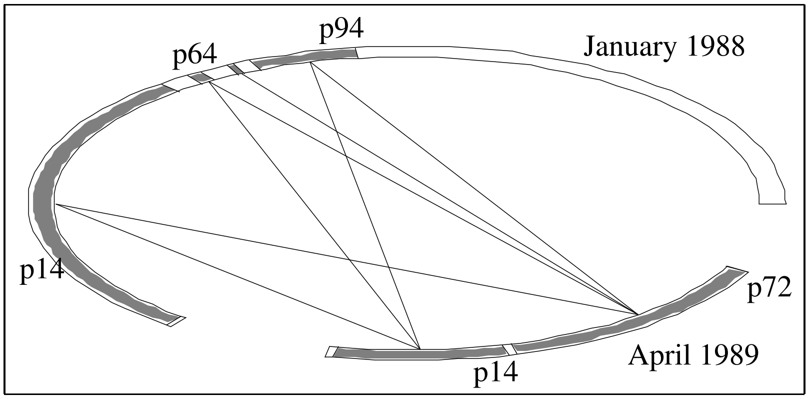


Figure 5.15: Comparison of cumulative lists

5.14 show the selected text for the January, 1988, and (the first half of the) April, 1989, versions of the document, respectively. Figure 5.15 shows the relationship between the documents. In this case there are multiple links displayed: those links were merged for the description since they were nearby, though not close enough to be merged.

- The text in all of 2205 *Cumulative List of Orphan Drug and Biological Designations, Jan 29, 1988* (p8-14,p18-72) is covered similarly in the similar passage of less significance in starting half of 23687 *Cumulative List of Orphan-Drug and Biological Designations, Feb 16,1989* (p6-12,p16-55,p57-85).
- The text in all of 23687 *Cumulative List of Orphan-Drug and Biological Designations, Feb 16,1989* (p6-12,p16-125,p127-221) is covered similarly in the similar passage of same relative significance in all of 29806 *Cumulative List of Orphan-Drug and Biological Designations, April 21, 1989* (p8-14,p18-227).
- The text in the final fifth of 6951 *Department of Defense, Supporting Agencies* (p11) is also discussed in the similar passage of same relative significance (but less coverage) in the starting quarter of 13105 *KGB* (p3).

5.2.5 Efficiency considerations

The link typing described above is effective—and, as Chapter 6 will show, reasonably accurate—at finding many types of links, but it is *not* efficient. The first part of the algorithm used in this study typing links between documents U and V is:

```
(1) Build part vectors $P_1$, $P_2$,...,$P_u$ for document $U$
(2) Build part vectors $R_1$, $R_2$,...,$R_v$ for document $V$
(3) As $i$ ranges from 1 to $u$
(4)     As $j$ ranges from 1 to $v$
(5)         Calculate sim($P_i$, $R_j$)
(6) Sort the similarities
```

The double loop of steps (3) and (4) clearly indicate that the algorithm is at least quadratic in the number of parts within a document. Steps (1) and (2), however, also require further processing so that the good/tenuous attribute can be determined, as discussed in Section 5.1.1 (page 50).

The sizes and numbers of paragraphs and sentences in documents range quite dramatically (in the Federal Register, the documents themselves range from 94 bytes to over 2.6Mb in length), so it is difficult to precisely describe the complexity of the algorithm. However, if we assume that documents have an average of p paragraphs, each containing n_p terms, then the complexity of building part vectors is:

$$O(p(n_p + n_p \log n_p))$$

That is, for each paragraph, the terms are identified in $O(n_p)$ time, and then they are sorted to create the vector.

Sentence vectors are needed for the good/tenuous analysis. If we assume each paragraph has s sentences on average, each with n_s sentences, then it will take $O(ps(n_s + n_s \log n_s))$ time to build the sentence vectors.

Whenever two paragraphs are compared, $O(n_p)$ time is needed to compute their similarity, and $O(s^2 n_s)$ is needed for the good/tenuous analysis (comparing every sentence within the two paragraphs). So the comparison takes time $O(n_p + s^2 n_s)$ per paragraph pair.

The algorithm's complexity, then, is

$$\begin{array}{lll} & O(p(n_p + n_p \log n_p)) & \text{(build part vectors)} \\ + & O(ps(n_s + n_s \log n_s)) & \text{(build subpart vectors)} \\ + & O(p^2(n_p + s^2 n_s)) & \text{(comparisons)} \\ + & O(p^2 \log p^2) & \text{(sort the results)} \end{array}$$

It is a simple matter to pre-build the part and subpart vectors for a collection, eliminating the first two components. The complexity then is:

$$O(p^2(n_p + s^2 n_s + \log p^2))$$

which for small documents is dominated by $O(p^2 s^2 n_s)$ and for larger documents with many more paragraphs, by $O(p^2 \log p^2)$.

The link *merging* process is quadratic in the number of links eligible for initiating a merge—*i.e.*, in the number of good and strong links. The algorithm is:

```
(1)  Find all link pair distances
(2)  While some distance is small enough
(3)       Merge closest two links
(4)       Recalculate incident distances
(5)  End-while
```

If g is the number of mergable links, then step (1) has complexity $O(g^2 + g \log g)$ to find the distances and sort them. Each pass through the loop of steps (2) through (5) reduces the number of links by one, so the loop is guaranteed to terminate. Step (3) is a constant time operation, and step (4) is linear in the number of links (not link pairs). So the overall complexity of the link merging loop is $O(g^2)$ meaning the complexity of the entire merging process is $O(g^2)$.

In the worst case, then, every link is a good and strong link, so $g = p$ and the complexity of the entire operation is dominated by the construction of the links.

Fortunately, the worst case rarely arises. The number of good and strong links is typically quite small compared to the total number of possible links. Further, when two links are merged, there are usually cascading merges that occur—if the endpoints of a link are merged, any other link pair incident on those endpoints must be merged, too. This event happens frequently in practice, which reduces the expected complexity.

Nonetheless, on a pair of very long documents (*e.g.*, 100 pages apiece), this algorithm cannot be used interactively. The implementation in the SMART system is designed for flexibility during experimentation, so some optimizations are possible. In general, however, this approach is probably best-suited for off-line processing.

5.3 Link type taxonomy

Once a comparison between two documents has been reduced to a simple graph and a few numbers—*e.g.*, convolution, link strengths—strikingly simple methods can be applied to type the links. Section 3.1.3 (page 24) presented a list of "automatic links" which were detectable using the techniques of this chapter. The following repeats the list and explains how each type of link can be recognized automatically.

Revision links are most noticeable because they have very little convolution: the linked paragraphs stay in roughly the same order. Because many of the paragraphs will be unchanged or only slightly modified, the similarities of the paragraph links must be extremely high (*e.g.*, over 0.80). The *Cumulative List of Orphan-Drug and Biological Designations* documents of Figure 5.3, for example, are detectable as revised documents.

(Note that without additional information, it is impossible to tell which document is the revised version of the other.)

Summary/expansion links are detectable by the many strong paragraph links but also by the many paragraphs which are *not* linked. Figure 5.16 shows the relationship between the paragraphs of two published papers. The one on the bottom of the figure is a conference paper. The one along the top is an expanded treatment of the same paper which was printed in a journal.

Summary/expansion links are very similarity to revision links, except that the amount of unlinked text must be different between the two documents. By detecting that the unlinked text lies exclusively at one end of one of the documents, the modified and broken documents of Figure 4.11 can be recognized for what they are, and not confused with an expanded treatment.

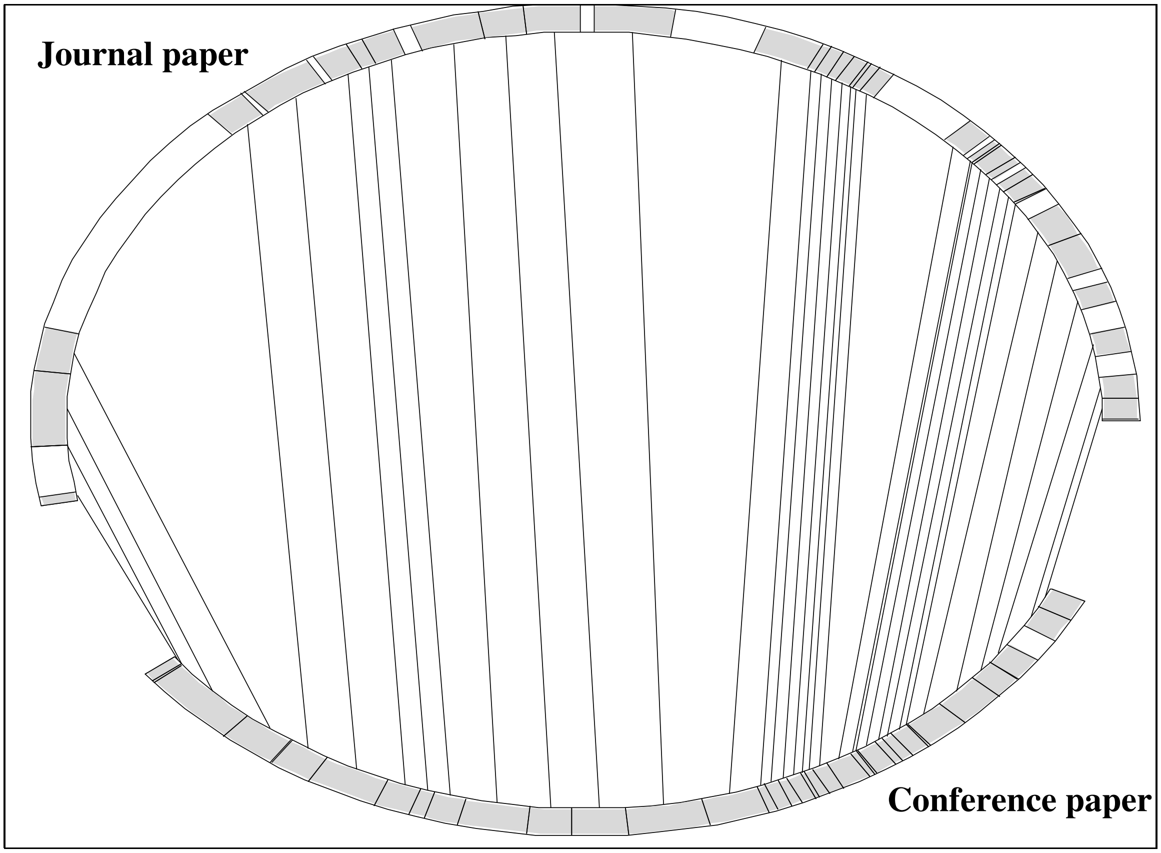


Figure 5.16: Conference and journal papers

The Kennedy and USA documents of Figure 4.14 are an extreme example of an expanded treatment (the similarities are not high enough to confuse it with a modified document broken into pieces). In that case, the expansion is so great and the similar material so focussed, that a better label for the link might be "expand to a much larger context."

Equivalence links are most other strongly-related pairs of documents. For example, the nuclear energy and nuclear weapons documents presented in Figure 4.13 (page 44) are "equivalent" in that they discuss highly-related topics.

Contrast links *could* be links between completely unrelated items (with zero similarity), but there is little value in such links. It is more useful to identify documents which have strong relationships within small sub-parts but are otherwise unrelated or only marginally related. The Chemistry and Physics example of Figure 4.12 is excellent illustration. The two documents can be linked as "contrasting" topics, though it may be more interesting to link the shared subtopic (molecular theory) and label the link "contrasting treatment."

Tangent links are "unusual" links. These include, for example, links between a pair of documents which fail the global/local restriction discussed in Section 2.3. It also includes links such as that between *Clouds* and *Georgia O'Keeffe* in Figure 4.9, where the O'Keeffe document is clearly related in a manner different from the other documents.

Comparison links are all documents linked because of similarity which do not fall into any of the other categories.

Aggregate links are groups of related documents. Aggregates are not discovered using the link analysis methods of this chapter. Section 5.4 below describes how aggregates are found.

To identify a link type, then, the following set of rules is used. We assume that the meta-links have been formed and we are considering exactly one meta-link. (Note that it is possible, therefore, for a pair of documents to have multiple relationships.) If the comparison of two *complete* documents (rather than their parts) is desired, then the meta-link can be expanded to include the entirety of both documents.

To type then link, first consider meta-links formed using *only* very strong links (those over 0.80).

- If the meta-link has a small amount of convolution (more than 60% of the links are skewed or parallel), label the meta-link "revision."

- If the proportion of unlinked text is substantially higher in one document than in the other, label the link from the first to the second document as "summary" and the reverse link as "expansion."
- Otherwise label the link "equivalence."

If there are no or very few extremely strong links, then consider meta-links created with all of the links:

- If there is little convolution, label the meta-link "equivalence."
- If there is a large amount of text in both endpoints that was not linked before the merging, label the meta-link "contrast."
- If the proportion of unlinked text is substantially higher in one document than in the other, label the link from the first to the second document as "condensed treatment" and the reverse link as "expanded treatment."
- Otherwise label the link "comparison."

"Tangent" and "aggregate" links are identified by other means.

5.4 Aggregation link type

Some link types do not require the complicated link merging procedure described above. One of the simplest link types to determine is an "aggregate" or "clump" link. An aggregate is a set of documents which are grouped together for a particular reason—typically for either structural or content reasons. (Note that because an aggregate includes many documents, it is not a simple uni-directional edge between two nodes of a hypertext.)

Structural aggregates might include all the documents which comprise a chapter or an entire book. These links are easily identified during automatic hypertext construction, since structural divisions of documents are typically annotated with a mark-up language or are found with simple pattern matching techniques (see the discussion of passage identification in Section 2.4 on page 15).

More interesting aggregates, however, are aggregates of documents which discuss similar material. The hypertext graph of Figure 2.1 (page 20) can be simplified substantially by collapsing associated nodes into an aggregate node. Figure 5.17 is identical to Figure 2.1 except that lines have been drawn separating it into components based upon the four documents most similar to *March music*.

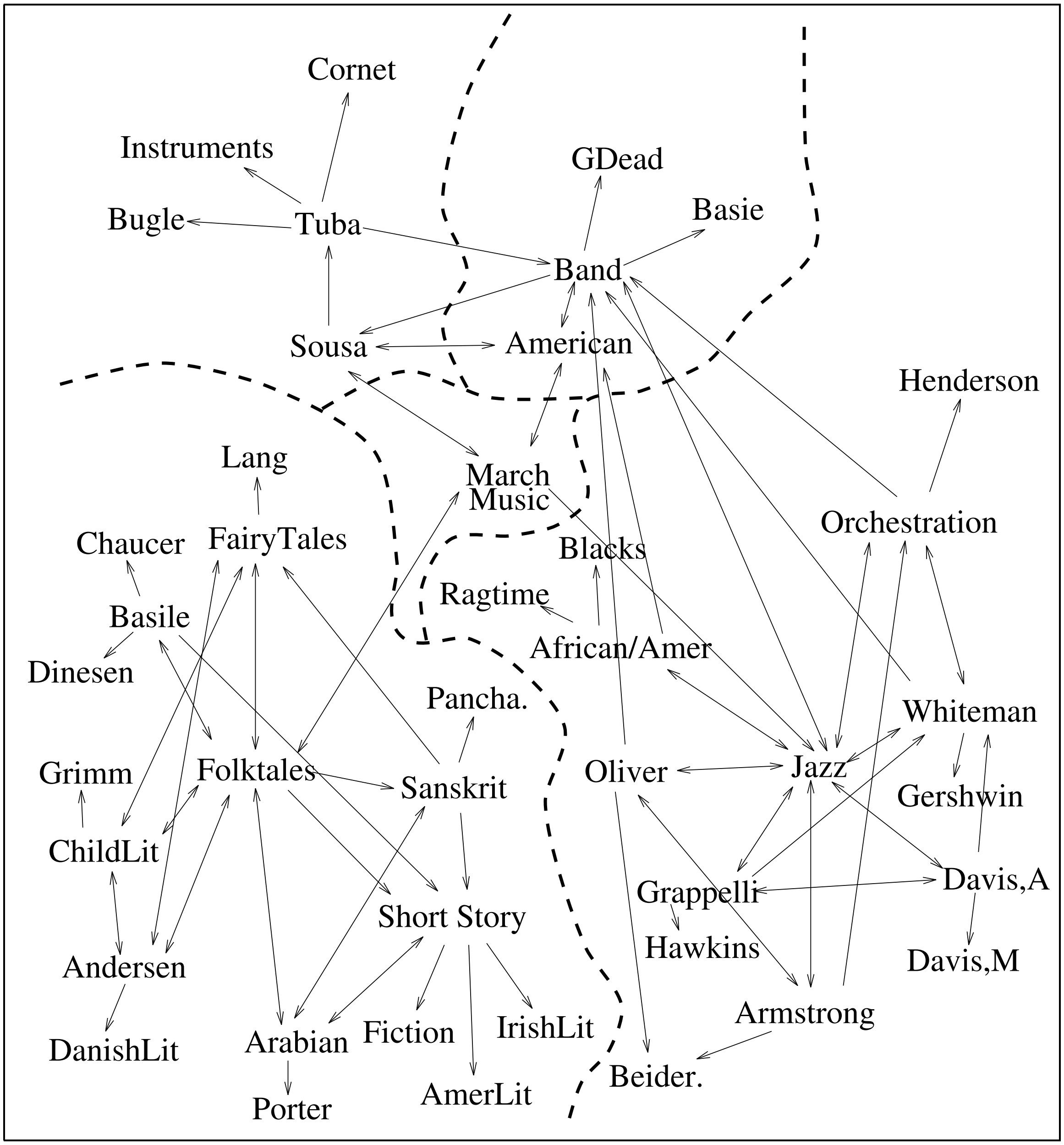


Figure 5.17: *March music* web with topic boundaries

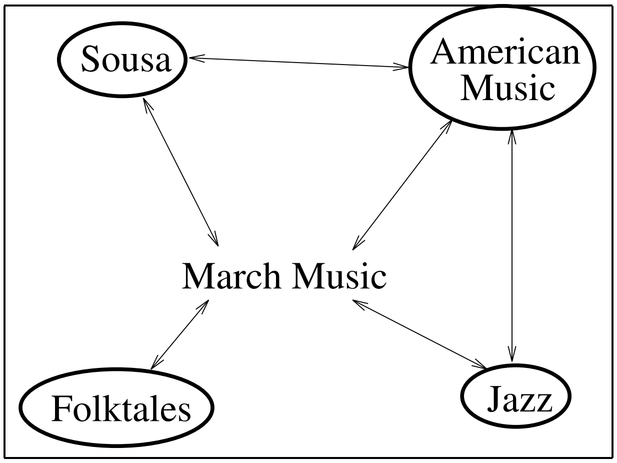


Figure 5.18: *March music* web collapsed

Figure 5.18 shows the far simpler hypertext which results from collapsing the grouped documents into a single aggregate: the aggregates are indicated by surrounding the node name with an oval. All of the linking information is still present, but the automatic aggregation makes the presentation much clearer.

Recent work has proposed techniques which could be used to provide a more accurate summary of the topics or themes which occur inside a set of aggregated documents.[SABS94] In this example, the aggregate is identified by the title of the document which was its starting point. (The graphs of the *March music* hypertext were manually created. Systems such as SaTellite[PT90] include algorithms for automatically doing two-dimensional layout of such graphs.)

Chapter 6

Evaluation

This chapter presents the results of an informal evaluation of the techniques discussed in Chapters 2 and 5. The intent of the evaluation was to determine whether the link typing methods were identifying link types which a user would find meaningful. There were substantial flaws in the evaluation, but some meaningful results came from it nonetheless. It demonstrated that some aspects of the link typing worked quite well, but that others need work.

6.1 Approach to evaluation

The basic approach to evaluating the link typing is to present a person with two texts, and have the evaluator answer questions about the relationship between those linked texts. The evaluators' answers can then be compared to computer-generated link types to see how well the automatic methods match human evaluations.

It would have been possible to evaluate the typed links themselves. That is, the evaluation set might have consisted of documents pairs and a link type. The evaluator would then be asked, for example, how well a particular link conformed to the notion of *Comparison*; however, such an approach requires that users have a consistent concept of "comparison" and is prone to too many interpretive errors.

For that reason, in this study, the evaluators were asked seven questions which rated narrower aspects of the relationship. This choice had the further advantage of allowing individual parts of the relationship-typing process to be evaluated separately. To that end, the SMART system was modified to determine automatically answers to several different questions and those answers were compared to the evaluators' answers.

The evaluation was designed, then, to find out:

1. How well did the part-comparison process identify and choose parts of docu-

ments over the whole document? Was the selected document part distinct from the rest of the document?

2. How well did the part decomposition and selection work? Did the selected part contain just one topic of discussion, or should it have been broken into even smaller parts?

3. How strong was the relationship between the two selected parts (the endpoints of the link)?

4. Is the expanded or condensed treatment of a topic detectable? In particular, does the "unlinked text" approach suggested in Section 5.3 actually identify those categories?

5. Does the amount of convolution in the meta-link actually correspond to anything? It can clearly be used to identify revision links. Can it be used to estimate how different are the writing styles of two documents?

Experience with the link typing suggested that the part selection techniques would work well, though not perfectly, and that it would only link texts with good relationships. This hypothesis has also been supported by research in Information Retrieval, since the linking approach used here is very similar to retrieval techniques.

The unlinked text, it was hypothesized, would be a good indicator of whether topic discussion was expanded or condensed.

Anecdotal evidence suggested that the amount of convolution would provide valuable information, but it was not clear exactly what it reflected.

The remainder of this chapter presents the evaluation itself. Section 6.2 describes how the evaluation was carried out, Section 6.2.1 describes the collection used for evaluation, and Section 6.3 is a detailed analysis of the results. The chapter concludes with a summary of the evaluation results in Section 6.4.

6.2 Evaluation method

The evaluation was arranged as follows. A random document from a collection was selected as a starting point. The eight documents most similar (and passing a local restriction test) to that random document were selected, creating a group of nine documents. The nine documents were then pairwise compared. For each of the 36 pairs, the following information was calculated and saved:

1. The most similar sub-parts of the two documents. In many cases, no pair of subparts had a higher similarity than the entire documents, so the documents were selected as a whole.

2. The amount of extraneous (unlinked) text which was added to create those parts during link merging (if an entire document was selected, this amount is zero);

3. The degree of link convolution within the meta-link joining the sub-parts. Note that if an entire document was selected, all links will be "wedge" links; if both entire documents are selected there is a single link, hence no convolution.

4. The sizes of the documents and their selected sub-parts.

Only the documents and their sub-parts were presented to the evaluators. The remaining information was saved to compare to the evaluators' answers.

The evaluators were presented with a pair of documents chosen randomly from the set of 36. The entire document was presented, but the most similar sub-parts (if any) were highlighted. The evaluation required answering seven questions which addressed the issues raised above.

Two questions dealt with the selected passages at the endpoints of the merged links:

- How distinct is the passage from its enclosing text? This question's purpose was to evaluate how well the part selection chose distinct sub-parts.

- How focussed is the material in the selected passage—does it cover multiple topics or a single topic?

When an entire document was selected, neither of those questions were meaningful, so they were not asked. The remaining questions dealt with the link itself:

- How close is the relationship between the linked passages?

- How does the topic coverage in the link source compare to the passage which is the link's sink? This was meant to evaluate the summary or expansion aspect of a relationship.

- How do the exposition styles of the link's source and sink compare? The hope was the convolution corresponded to "writing style" or "exposition style."

In the discussion of evaluation results in Section 6.3, each of the evaluation questions is presented in full.

6.2.1 Evaluation collection

For these evaluations, three sets of nine documents were chosen. The starting documents were:

Table 6.1: *Espionage* evaluation group

Docid	Sim	Size (bytes)	Title and Description
8316	1.00	20286	*Espionage* or spying. Describes espionage itself, spying organizations, justification, techniques, history, industrial and political aspects, and implications
4868	0.52	5889	*CIA*. Gives the history and activities of the agency. Discusses surrounding controversy and investigations.
22285	0.35	1610	*Lewis Terman* devised the notion of IQ and one of the tests to measure it.
8317	0.29	2207	*Espionage Act of 1917* is US legislation which describes penalties for spying against the US. This legislation indirectly prompted the "clear and present danger" test.
13105	0.28	2864	*KGB* covers the functions and a brief history of the Soviet agency.
6951	0.27	6116	*Defense department* discusses the role of this executive department in the US government. It covers the command structure and lists agencies which support it.
12157	0.27	1821	*Intelligence* is definition of "intelligence" which spans several paragraphs.
7616	0.25	1057	*Allen Walsh Dulles* was CIA directory from 1953 to 1961.
8680	0.25	7426	*FBI* covers the history and jurisdiction of the agency. It also includes a lengthy description of the activities of the FBI.

1. An encyclopedia article entitled *Espionage* which discusses spying, its justification, its techniques, its history, and its effects. Table 6.1 summarizes the nine documents in this set.

2. An encyclopedia article on William Shakespeare. The nine documents are summarized in Table 6.2.

3. A group of documents taken from the Federal Register. TREC query 10[Har92b] was run against the collection, and 9 of the top-matching documents were selected for the evaluation. Query 10 is:

   > To be relevant, a document must include a reference to at least one specific potential Acquired Immune Deficiency Syndrome (AIDS) or AIDS Related Complex treatment.

   The information about the documents is presented in Table 6.3. In this set, the similarity in the table is between the document and the query.

6.3 Evaluation results

This section describes the results of the evaluation.

6.3.1 General results

Eleven individuals participated in the evaluation. They were all computer scientists by training. Their participation was voluntary with no compensation. The evaluation task was not monitored beyond acquiring the answers. Only 2 evaluators evaluated all 108 (thrice 36) pairs. The breakdown is presented in Table 6.4. Evaluations for one document pair were discarded because of an error during its processing that was not detected until the evaluation was complete.

An interesting issue is the consistency of the evaluators: for the same document pair, how well did the answers coincide? 107 document pairs were evaluated, with 7 questions per pair. That means that there were 749 questions which were answered. How well did the evaluators agree on those 749 questions?

168 of those questions were actually never asked, since they dealt with parts of documents when the entire document was chosen. For an additional 102 (13.6%) of the 749 questions, all evaluators gave the same answer.

The remaining 479 (64.0%) of the questions had evaluations which were inconsistent, though far-ranging inconsistency was rare. Table 6.5 indicates how far apart the answers were for the various questions. For example, if a question had evaluations answers of only 2 and 3, it would have a spread of 1. For almost all of the questions,

Table 6.2: *Shakespeare* evaluation group

Docid	Sim	Size (bytes)	Title and Description
20986	1.00	17942	*William Shakespeare* discusses the playwright's life, his plays, and his reputation.
8176	0.46	25331	*English Literature* covers literature produced in England from the 5th century to the present.
22301	0.43	1651	*Dame Ellen Alicia Terry* was the "leading lady of the English stage" around the turn of the century; she was famous in particular for her roles in Shakespeare productions.
15069	0.34	3114	*Christopher Marlowe* was the most famous English playwright before Shakespeare.
7514	0.33	69357	*Drama and Dramatic Arts* describes how drama is portrayed in theatrical performances, and describes the history of the dramatic arts from Greek theatre through contemporary theatre. It covers both Western and Eastern forms of drama, though it concentrates far more on the former.
14577	0.33	1460	*Macbeth* was the king of Scotland from 1040 through 1057. He is also a character in Shakespeare's tragedy *Macbeth*.
13030	0.33	2604	*Kemble* describes the Kemble family of English actors, two of whom were particularly noted for performances in Shakespeare plays.
13569	0.31	2005	*Charles Lamb* was an English essayist who wrote the popular *Lamb's Tales from Shakespeare*.
13465	0.29	977	*Thomas Kyd* was one of the "most important dramatists of the early Elizabethan period." Shakespeare imitated Kyd's use of shocking situations.

Table 6.3: *AIDS* evaluation group

Docid	Sim	Size (bytes)	Title and Description
22221	0.59	17685	*Vet Admin, final rule, Systemic diseases (1/30/89)* discusses how the Veterans Administration has amended the list of disabilities to include several AIDS-related symptoms.
26911	0.39	4100	*FDA notice, Antiviral drugs (3/22/89)* is a notice from the FDA requesting nominations for members to serve on an antiviral drugs committee.
21848	0.38	1090	*HRS, AIDS demo projects extension (1/24/89)* is a short document mentioning that an application due date has been extended.
23687	0.38	60719	*FDA, Cumulative list of orphan drugs and biological designations (2/16/89)* is a list of drugs for which the sponsor has exclusive manufacturing rights. (The document was judged relevant to query 10.)
29806	0.38	63352	*FDA, Cumulative list of orphan drugs and biological designations (4/21/89)* is a later version of the previous document. (This document was also judged relevant.)
25238	0.37	12672	*Public Health Service, new system of records (3/6/89)* describes a proposal to establish a new Privacy Act system of records.
28575	0.37	2530	*HRS, AIDS drug reimbursement program (4/10/89)* describes an allocation of $5 million given to states to help them with costs of providing AZT. (This document was judged relevant to the query.)
21886	0.36	17795	*Table of contents (1/24/89)* summarizes in a sentence or phrase each of several entries in the Federal Register, including some related to AIDS.
2205	0.32	20198	*FDA, Cumulative list of orphan drugs and biological designations (1/29/88)* is the *first half* of an earlier version of the documents listed above. (The second half of this document was judged relevant.)

Table 6.4: Breakdown of evaluations

	Evaluators	Number of Evaluations			
Group	All/Total	Total	Per pair		
			Min	Max	Avg
Espionage	4/5	147	4	5	4.1
Shakespeare	3/7	140	3	5	4.0
AIDS	3/3	108	3	3	3.0
Total		395	3	6	3.7

Table 6.5: Range of answers to same question

Max	Count	Prop. of 479	
spread			Summed
1	197	41.1%	—
2	135	28.2%	69.3%
3	123	25.7%	95.0%
4	8	1.7%	96.7%
5	16	3.3%	100.0%

Table 6.6: "Distinctiveness" breakdown

Num	Description	Evaluations	
0	n/a	30	(6.61%)
1	very	137	(30.18%)
2	somewhat	185	(40.75%)
3	badly	102	(22.47%)
	TOTAL	454	

the evaluators strayed no more than 2 from each other when answering. The spread clearly indicates some variability in answering, but not enough to cause concern.

6.3.2 Distinctiveness

The first two of the seven questions were an effort to evaluate how effectively the link-typing chose meaningful portions of the two documents. For both the source and the destination of the link, the user was asked:

> How distinct is the highlighted text within the source (destination) document?
>
> 0. *n/a*: the entire document was linked;
> 1. *Very*: highlighted topic is clearly distinct from rest of document;
> 2. *Somewhat*;
> 3. *Badly*: there is nothing distinguished about the highlighted text.

107 links were evaluated, so 214 link endpoints were considered. For 84 of those endpoints, the system did not select a document passage, instead linking the entire document. In those cases, the evaluators were not given a chance to answer the distinctiveness questions; a value of "0" was assumed.

The 395 evaluations resulted in 790 distinctiveness ratings. 336 of those corresponded to the 84 cases where the entire document was linked. The remaining evaluations break down as shown in Table 6.6. In less than a quarter of the cases, the evaluators felt that the selected text was not distinct within its containing document. It is also interesting to note that 30 evaluations rated the entire document as linked (the "n/a" answer), even though the system did not.

It was hoped that the distinctiveness would diminish as the size of the selected passage grew in comparison to its containing document. Unfortunately, Table 6.7 shows that the evaluations do not particularly support that. Although it is true

Table 6.7: Distinctiveness vs. passage size

Size of passage	n/a	Number of evaluations very	some	bad
0–9%	3	61	85	43
10–19%	0	5	16	5
20–29%	0	23	33	11
30–39%	0	12	22	6
40–49%	1	16	9	13
50–59%	0	7	12	11
60–69%	0	8	4	0
70–79%	0	0	1	2
80–89%	18	0	0	5
90–99%	8	5	3	6
100%	0	0	0	0

that relatively fewer small links are rated “badly distinctive” (most small passages are rated at least somewhat distinct), the passages which were rated “bad” range throughout the possible sizes. Note that most of the 30 “not applicable” ratings occur when the link size is quite large: some evaluators evidently felt that most of the document was close enough to “all of the document” to warrant such an answer.

Although some badly distinctive text is selected, over 2/3 of the time very or somewhat distinctive text is chosen. These results indicate that the part breakdown and comparisons successfully find portions of text which are distinct from the text as a whole. Our experience suggested that automatically selecting text portions was useful; these statistics support our intuition.

6.3.3 Focus

The “distinctiveness” questions were an attempt to determine if the selected text portion was distinct from the text in which it was embedded. The next two questions asked whether the selected text included a single or multiple topics (or subtopics) of discussion. That is, should the passage have been broken into an even smaller unit. The specific question was:

Table 6.8: "Focus" breakdown

Num	Description	Evaluations	
0	n/a	13	(2.86%)
1	tightly	184	(40.53%)
2	somewhat	201	(44.27%)
3	loosely	56	(12.33%)
	TOTAL	454	

How focussed is the highlighted text within the source (destination) document?

0. *n/a*: the entire document was linked;
1. *Tightly*: only one topic is covered within the text;
2. *Somewhat*: multiple subtopics, but they are all related;
3. *Loosely*: the text includes several topics with no relationship.

As with the "distinctive" questions, if an entire document was selected as the link endpoint, the focus of the chosen material was considered unimportant: the issue is how well focussed is a selected passage, not an entire document. Of the 790 evaluations of link endpoints, only 454 passages are considered. The evaluations are summarized in Table 6.8. The table indicates that the users believed the textual discussion was tightly or somewhat focussed in almost all cases. There is little relationship between the evaluation of a passage's "focus" and the passage's size relative to the size of the document, although Table 6.9 shows there is a slight tendency for larger passages to be less well focussed.

It seems probable that the relative size of the link may be a less useful predictor of focus than the absolute size, but oddly enough Table 6.10 shows that the "badly focussed" documents are distributed throughout the range of passage size and, in fact, more concentrated when the passage sizes are smaller.

The "focus" of the document part is used by automatic link typing to look for summary and expansion links. It is thus not just important that the endpoint be focussed, but also that the system be able to judge the degree of focus. To that end, a "correct" answer to the evaluation question was automatically generated for every case according to the amount of extraneous text that was added during link merging (see Section 5.2.2). The thresholds used are presented in Table 6.11. For this set of evaluations, the link merging cutoffs were set low enough that no more than 20% extraneous text was likely to occur, though there were a few cases where almost 75%

Table 6.9: Focus vs. passage size

Size of endpoint	n/a	Number of evaluations tight	some	loose
0–9%	0	93	85	16
10–19%	0	9	11	6
20–29%	0	30	32	5
30–39%	0	9	22	9
40–49%	1	15	18	5
50–59%	0	13	10	7
60–69%	0	9	3	0
70–79%	0	0	2	1
80–89%	8	5	7	1
90–99%	4	1	11	6
100%	0	0	0	0
TOTAL	13	184	201	56

Table 6.10: Focus vs. absolute size

Size of endpoint	n/a	Number of evaluations tight	some	loose
0– 1000 bytes	6	133	94	12
1– 2000 bytes	1	36	61	11
2– 3000 bytes	2	8	18	10
3– 4000 bytes	0	7	7	1
4– 5000 bytes	0	0	10	7
5– 6000 bytes	2	0	1	1
6– 7000 bytes	0	0	2	2
7– 8000 bytes	0	0	2	1
11–12000 bytes	0	0	0	4
16–17000 bytes	0	0	3	3
23–24000 bytes	0	0	1	2
50–51000 bytes	1	0	1	1
51–52000 bytes	1	0	1	1
TOTAL	13	184	201	56

Table 6.11: Threshold for defining "focus"

Percent extra	Choice
10.1–100%	Loosely focussed
4.1–10%	Somewhat focussed
0.0–4%	Tightly focussed

Table 6.12: System vs. evaluators on focus

	Evaluator judgements							
System	n/a		tight		some		loose	
tight	10	(2.20%)	179	(39.43%)	182	(40.09%)	34	(7.49%)
some	3	(0.66%)	0	(0.00%)	4	(0.88%)	3	(0.66%)
loose	0	(0.00%)	5	(1.10%)	15	(3.30%)	19	(4.19%)

of the linked material was considered extraneous.

A comparison between the evaluators and the system is presented in the Table 6.12. The table suggests that the algorithm is quite good at identifying "tightly" focussed links, but has great difficulty distinguishing them from "somewhat" focussed links.

This evaluation parameter clearly indicates that more work needs to be done on identifying the focus of selected material. It is likely that the analysis needs to be local to the material, and not as a result of comparing it to an external document or document part. The recent successes in theme extraction and document summarization further suggest such an approach.[SS94]

6.3.4 Relatedness

Because of the way the evaluation documents were chosen, there was likely to be some similarity between any two documents in the same set—they were all believed similar to a particular starting document. However, the relationship between them can range from extremely strong, through vague, to (surprisingly) none at all. Given two document parts that were linked, the evaluators were asked the following question:

Table 6.13: Relatedness assignment criteria

Type	Reason for assigning
No relation	The most highly similar parts had a similarity below 0.10.
Slightly	Not "no relation," but more than 30% of either the source or sink of the link was extraneous text.
Related	None of the other cases applied.
Strongly	The average of the "strong" links was over 0.60 (but below 0.80)
or	The average of the "strong" links was over 0.40 and the more than a third of the links were "strong" (but not enough to make an "identical" rating).
Identical	The average of the "strong" links was a similarity over 0.80
or	The average of the "strong" links was over 0.50 and the over half of the links were "strong."

How related is the material in the links?

1. *Not at all*: why is there a link?
2. *Slightly*: vague relationship
3. *Related*: good similarity, but topics are not the same
4. *Strongly*: very similar topics are covered
5. *Identical*: almost exactly the same material is presented

The Smart system was used to assign automatically one of those categories to each of the 107 links based upon the criteria in Table 6.13. The links were assigned to categories are listed in Table 6.14.

395 links were evaluated for their relatedness. The evaluators judgements break down as presented in Table 6.15. (It is unclear what the judgement "n/a" meant—*i.e.*, why it is any different from "No relation". Those 6 responses are omitted from the analysis below.)

Overall, the evaluators and the system were somewhat consistent, though there is substantial disagreement in the number of links assigned to "slightly related" and "related." Table 6.16 does a breakdown of how the evaluators and the system compared for each possible rating. This table indicates that the system judgements were

Table 6.14: Assignment of relatedness

Type	Count	
No relation	45	(42.06%)
Slight	6	(5.61%)
Related	50	(46.73%)
Strong	5	(4.67%)
Identical	1	(0.93%)
TOTAL	107	

Table 6.15: Evaluations of relatedness

Type	Count	
n/a	6	(1.52%)
No relation	169	(42.78%)
Slight	124	(31.39%)
Related	61	(15.44%)
Strong	26	(6.58%)
Identical	9	(2.28%)
TOTAL	395	

Table 6.16: System vs. evaluators on relatedness

System	Evaluator choose this									
Choice	None		Slight		Related		Strong		Ident	
none	85	21.85%	59	15.17%	23	5.91%	7	1.80%	0	0.00%
slight	7	1.80%	9	2.31%	5	1.29%	1	0.26%	0	0.00%
relate	72	18.51%	51	13.11%	32	8.23%	17	4.37%	3	0.77%
strong	4	1.03%	4	1.03%	1	0.26%	0	0.00%	6	1.54%
ident	1	0.26%	1	0.26%	0	0.00%	1	0.26%	0	0.00%

Table 6.17: Coverage assignment criteria

Type	Reason for choosing type
Similar	Source and sink have extraneous text within 10% of each other.
More/less	The difference in extraneous text is between 10 and 40%.
Much more/less	The difference is more than 40%.

acceptable at the low extreme, inaccurate in the middle, and inconclusive at the high extreme. In the cases where the system judgement was "No relation," the evaluators' responses were skewed to the low end—more than 80% (144 of 174) of the judgements were for no or slight relationship.

Unfortunately, when Smart judged documents "related," the evaluators tended to disagree. Although almost 60% of the judgements were in the "slightly related" to "strongly related" range, they were clearly skewed heavily toward "no relation." This problem may stem from the same difficulties which prevent an information retrieval system from achieving perfect precision: some documents with high similarity will not be relevant.

6.3.5 Topic coverage

The users were asked to evaluate the coverage of the linked texts:

> How is the topic coverage in the destination versus the source?
>
> 0. *totally different*;
> 1. *Much more*: extra subtopics of source omitted in destination;
> 2. *More*;
> 3. *Similar*: equal mix of extraneous subtopics;
> 4. *Less*;
> 5. *Much less*: additional subtopics introduced in destination.

The intent was that the system would compare the extraneous material merged into the texts in order to decide how well the topics balanced. The system assigned each link to a category as listed in Table 6.17. The "extraneous" measure is the percent of the linked passage which was added during link merging but which does not contribute to the similarity between the link's endpoints. In this set of documents, there was

Table 6.18: Evaluations of coverage

Answer	Count	
0	228	(57.72%)
1	14	(3.54%)
2	7	(1.77%)
3	105	(26.58%)
4	14	(3.54%)
5	27	(6.84%)
Total	395	

Table 6.19: System-assigned coverage breakdown

Answer	Count	
0	0	(0.00%)
1	1	(0.93%)
2	4	(3.74%)
3	97	(90.65%)
4	3	(2.80%)
5	2	(1.87%)
Total	107	

rarely much extraneous text–only 10% of the link endpoints actually included any unrelated text. The percent ranged from 0.88% to 77.34%, but the small numbers of cases made this difficult to evaluate.

The breakdown of the user's evaluations is presented in Table 6.18, the distribution of the system-generated answers is in Table 6.19, and the correspondence between the system and evaluator answers is shown in Table 6.20. If the "totally different" case (0) is disregarded, the system does reasonably well at identifying the cases of "similar" topic coverage (the item where both system and evaluators chose "3"). However, it tends to misidentify most texts as having "similar" coverage, clearly suggesting that extraneous text is not sufficient for making this judgement.

In 108 of the 228 cases where the evaluators indicated "totally different," the system had detected only a "slight relation." Using the relationship similarity as a filter to detect unrelated pairs would have been helpful.

Table 6.20: System vs. evaluators on coverage

System's	Evaluator choose this					
Choice	0	1	2	3	4	5
0	0	0	0	0	0	0
1	1	0	0	0	1	1
2	10	2	0	3	0	0
3	207	11	7	97	13	24
4	5	1	0	4	0	1
5	5	0	0	1	0	1

Table 6.21: Breakdown of exposition evaluations

Rating	Count	
No relation	198	(50.13%)
Identical	33	(8.35%)
Similar	67	(16.96%)
Scrambled	97	(24.56%)
Total	395	

6.3.6 Exposition style

The evaluators were asked to provide an estimation of how differently the two endpoints of a link presented the text:

> How do the exposition styles of the two links compare?
>
> 0. *no relation between links*;
> 1. *Identical*: essentially the same order of description;
> 2. *Similar*;
> 3. *Scrambled*: completely different way of discussing material.

Unfortunately, it turns out that this is particularly difficult task to explain, and not much easier to carry out. The evaluators' judgements break down as shown in Table 6.21. The hope was that the link "convolution" (see Section 5.2.1) would be a reasonable predictor of this value; unfortunately, that hope appears to be poorly founded. The system answers were calculated as shown in Table 6.22. Table 6.23 shows that the system-generated judgements appear to be similar to the evaluators',

Table 6.22: Exposition style criteria

Type	Reason for choosing type
No relation	No links between parts
Identical	More than 80% of the merged links were parallel or non-crossing
Similar	Between 50 and 80% of the links were parallel or non-crossing
Scrambled	All other cases.

Table 6.23: System assignment of exposition style

Answer	Count	
0	80	(74.77%)
1	1	(0.93%)
2	9	(8.41%)
3	17	(15.89%)
Total	107	

Table 6.24: System vs. evaluators on exposition style

System's	Evaluator choose this			
Choice	0	1	2	3
0	151	23	55	72
1	0	3	0	0
2	15	6	4	4
3	32	1	8	21

but Table 6.24 shows that the overlap between the two is unfortunately very bad. The uncertain meaning of "exposition style" makes the value of this statistic minimal.

6.4 Overall conclusions

The informal evaluation described in this chapter was flawed in numerous ways, so only tentative conclusions can be drawn. It appears that:

- Chosen passages are usually distinct from their enclosing text. This confirms the hypothesis that the part comparison process does a good job of identifying and choosing a part from the whole document.
- The material in the chosen passages is typically well focussed, supporting the hypothesis that one topic of discussion would be found in selected document subparts. However, the measure of extraneous text is not sufficient for determining how well focussed the material is.
- It is as difficult to determine the "relatedness" of two pieces of text as it is to be certain of the relevance of a document to a query.
- A comparison of extraneous text works to a limited degree in comparing the coverage of two texts, but it can only be used if the texts are known to be similar. The evaluation suggests that the hypothesis is correct, but does not allow a definite conclusion.
- No conclusion can be drawn about the utility of a "convolution" measure—and in particularly how that relates to the exposition styles of two documents.

The conclusions of this evaluation are disappointingly inconclusive. However, anecdotal evidence such as that presented by the examples in Chapter 5 suggests that the measures used are valuable. At a minimum, the evaluation failed to refute the hypotheses.

It remains to be discovered whether the evaluation was completely flawed, whether the measures are metrics for unevaluated relationships, or—most likely—whether some combination of the two occurred.

Chapter 7

Conclusion and Future Work

In this thesis, we have presented methods for automatically linking related documents, showed two novel visualization techniques, and have described a process for automatically assigning a type to document relationships. In Section 7.1 of this final chapter, we review the major contributions of the thesis; Section 7.2 discusses addition research that should be done in the future.

7.1 Contributions

In a discussion of extracting "implicit inter-document links," Glushko concludes:

> When we first began working in hypertext several years ago, we expected that it would soon be possible to extract these implicit links automatically with natural language processing or clever indexing techniques..., but we have been disappointed so far and we are starting to conclude that implicit [inter-]document links are best identified by the hypertext reader.[Glu89]

This thesis has demonstrated that although some types of implicit relationships between documents are not yet within our grasp–*i.e.*, the "manual" links of Section 3.1.2—Glushko's judgement was premature. Using no natural language processing, and standard indexing techniques from Information Retrieval, we have shown methods for locating and describing inter-document links.

The major contributions of this thesis were the following:

- Adjusting information retrieval techniques for use in developing high-quality document and document passage links. (The work in this area is not new to this thesis, though it has rarely been used in this manner.)

- Development of a uniform visualization technique for showing the relationship between documents more clearly than a textual description. The uniform visualization is useful for identifying patterns of relationships and in particular for locating documents within a group which have an unusual relationship to the group.

- Development of a variable visualization method. This method is much like the uniform visualization, but it also implicitly includes document and passage size information. This visualization provides an excellent sense of the relationship between two (or more) documents.

- Classification techniques for automatically typing the relationship between two (or more) documents. Automatic typing is a necessary component of any automatic hypertext system: without link types, the "lost in hyperspace" problem is almost unavoidable.

- Preliminary evaluation of automatic typing. Although the informal evaluation presented in this thesis was flawed, it did provide useful information about how additional evaluations should be done. Despite its problems, the evaluation also suggested that portions of the automatic typing work very well.

This thesis clearly demonstrated that automatic hypertext construction is a viable option. Indeed, for the rapidly growing and dynamically adjusting databases available on modern computers, automatic hypertext linking may be the *only* viable option.

7.2 Future work

This thesis has laid the groundwork for automatic hypertext construction. Additional research is necessary to determine the value of some link types and to investigate other options for linking. The following list details some possible directions that this research can be taken.

- Perform a more rigorous evaluation of the link typing. Such an evaluation almost certainly requires performing typing on a larger and more varied collection of documents. The encyclopedia used in this thesis suffers the disadvantage of an editorial policy which necessarily constrained the writing style of the article. The Federal Register lacks that limitation, but does not contain very interesting documents.

- Extend this work to more heterogeneous text collections. By adding documents from different areas (*e.g.*, *Wall Street Journal* articles or historical documents such as *The Federalist Papers* and the *Articles of Confederation*) we can produce links within a particular subject area and which cross the sub-collection boundaries.

- Use relevance feedback approaches to limit the spread of an automatically constructed hypertext. The hypertext web created starting from a particular document (such as the *March music* example in Figure 2.1 on page 20) can easily drift from the starting topic—as topics are more links from the starting document, they will be less and less like the starting document.

  Relevance feedback is a technique for adjusting a query to be more like relevant documents.[Sal89] It might be useful to build a hypertext web where the linked documents are constrained to be more like the starting query. For example, *Jazz* is an article similar to *March music*. Rather than proceeding by looking for documents similar to *Jazz*, the system would look for documents similar to *Jazz* with *March music* as a known relevant document. This approach should keep the retrieved documents from drifting too far from the original topic.

  Note that these "feedback groups" are merely another type of aggregate group. They can co-exist with the "non-feedback groups" since each serves a different purpose.

- Attempt to identify the kinds of information that are useful for developing more and more accurate types of links. At the moment, links are made based upon a simple (but highly effective) similarity measure. What other (general; not domain-specific) information would be useful to find and type links?

- Pursue visualization techniques. The graphs presented in Chapter 4 were the inspiration for much of the link typing described in this thesis. Other types of visualization will no doubt highlight different relationship—hence, different link types—between documents.

Appendix A

Similarity merging

This appendix describes the type of similarity combination used in the link merging of Chapter 5. This material is presented in an appendix because although it is interesting, it is of tangential rather than central importance to the rest of the thesis.

Section A.1 discusses necessary requirements for any linking scheme. That is followed by a presentation of the merging formula used in this work, and Section A.4 includes a theorem and proof that the formula satisfies the needs.

A.1 Merging basics

When a pair of links is merged, the resulting "meta-link" will be between different text (*i.e.*, one or both of the endpoints will be expanded), so the link will almost definitely represent a different similarity than it did before. It would be possible to recompute the similarity by calculating the vectors for the merged endpoints and computing the similarity, but it is possible to achieve a similar effect with less computation by combining the similarities appropriately.

A method for combining similarities is acceptable only if it behaves in a reasonable and predictable manner. The following discussion lists 7 properties that must be satisfied when several links are combined into a single meta-link. To help illustrate the properties, abstract representations of link pairs like those in Figure A.1 are used. In Figure A.1, the link pair in question is represented by lines labelled A and B. A is the strength of a link between endpoints of size α and ϵ; B is the strength of a link with endpoints having sizes β and π. The endpoints at the top of the figure (with sizes α and β) both come from the same document, as do the pair with sizes ϵ and π.

Figure A.1 also includes two additional links. These are needed because the similarities other than A and B may also be known. In this example, there is no similarity between the endpoints labelled ϵ and β, so it is drawn as a dotted line and labelled

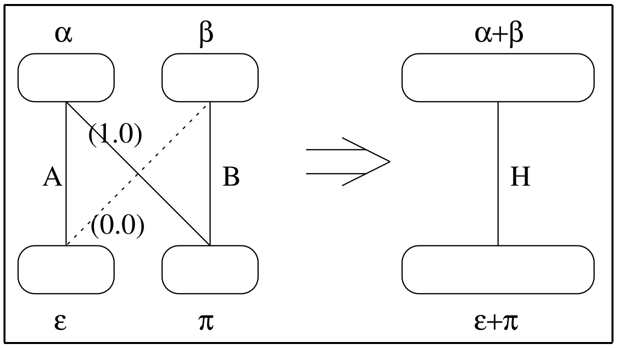


Figure A.1: Sample links for merging

"0.0". Similarly, the edge labelled "1.0" represents a perfect similarity between the endpoints (recall that normalized similarities are being used, so they range from 0.0 to 1.0).

The right-hand portion of the sample, to the right of the arrow, is intended to represent the result of merging the two links together. The text portions of sizes α and β from the top document are merged to create a larger portion of size $\alpha + \beta$. The bottom endpoints are merged to create a meta-endpoint of size $\epsilon + \pi$. The meta-link has a similarity of H, which is somehow based upon the original similarities and endpoint sizes.

Note that all of the link merging process discussed in Chapter 5 can be represented in this manner. When links are merged with adjacent endpoints, the match is clear. Consider, though, a wedge link such as in Figure 5.4b (page 54). In that case, there is unlinked text which must be merged into the meta-link at the same time. Such a wedge link is merged together as a two-step process: first the unlinked passage is merged into one of the links, pretending that there is a 0.0-weighted edge between the two drawn edges; second the new meta-link is merged into the other link.

A.2 Merging requirements

Consider what properties are required of a link merging technique—*i.e.*, what relationship must the meta-link's similarity bear to the original links' similarities? A similarity combination method is useful only if it satisfies each of the following properties:

1. **Extraneous reduces.** If two portions of text are linked, and one of the portions is expanded by added non-linked text (with similarity of 0.0), the weight of the

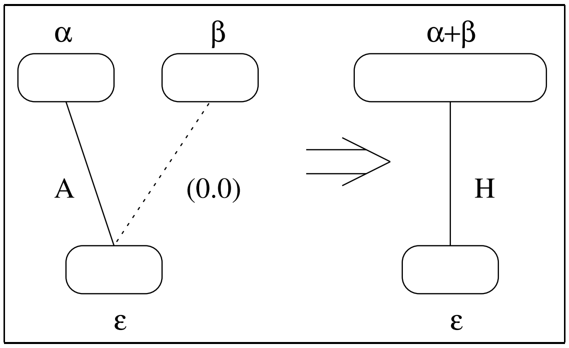


Figure A.2: Expanding with extraneous text

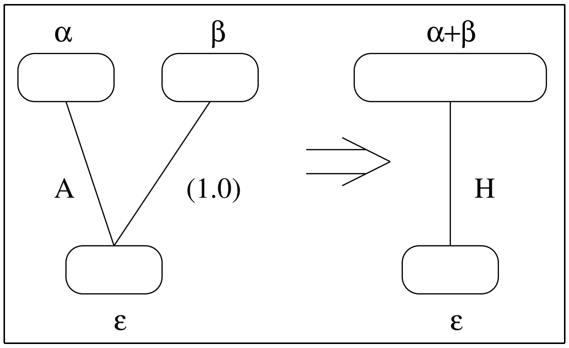


Figure A.3: Expanding with identical text

link must decrease. In Figure A.2, that means that $H \leq A$. Further, if $\beta = 0$ then $H = A$ (since the link was not changed), and if $\beta > 0$ then $H < A$. Finally, as the amount of extraneous text added grows, the link strength should drop correspondingly—*i.e.*, as $\beta \to \infty$, $H \to 0$.

2. **Identical improves.** If two portions of text are linked, and one of the portions is expanded by added perfectly-similar text, the weight of the link must not decrease—in fact, it must increase if the original link was not an identity link. This case is illustrated in Figure A.3 where if $\beta > 0$, we require that $H \geq A$ and if $A < 1$ also, then $H > A$. Further, as $\beta \to \infty$, $H \to 1$. If $\beta = 0$, it is preferable that $H = A$, though that case is nonsensical unless $\epsilon = 0$ also: the parts cannot have perfect similarity if only one of them is zero-length.

3. **Partial idempotence.** If an endpoint of a link is "merged" with itself, the "combination" must be identical to the original link. That is, in Figure A.4,

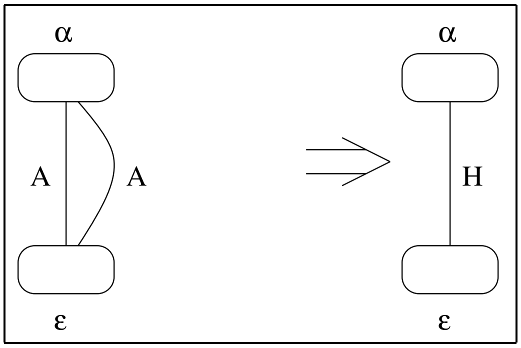


Figure A.4: Partial idempotence of merge

$H = A$.

4. **Commutative** It cannot matter whether link A is merged into link B or whether link B is merged into link A.

5. **Symmetry** When a non-wedge pair of links is merged, it cannot matter which document parts are merged first. Figure A.5 shows a merge proceeding in two different ways: along the top, the top document parts are merged first; along the bottom, the bottom parts are merged first. We require that $G = H$.

6. **Wedge associative** When three links share a common endpoint and are merged together, it cannot matter which of the three is merged first. Figure A.6 shows this case, where along the top, links A and B are merged and then link C is merged into the result; along the bottom, links B and C are first merged and then link A is merged in. We require that $H = G$.

7. **Associative** If three links are all merged together into a larger meta link, they will be merged as a two-step process. First one pair of links is merged, and then the third link is merged into that meta-link. In this case, it must not matter which pair is merged first. That is, in Figure A.7, it cannot matter whether A and B are merged first (as at the top of the figure), or whether B and C are merged first: in either case, once the next merge is completed, it must be that $G = H$. (Additional links which may or may not be present are drawn as dotted lines.)

Any formula for deriving a new similarity for a meta-link must satisfy all of the properties listed above, or its value is nil. We claim those properties as necessary, but are not confident that they are sufficient.

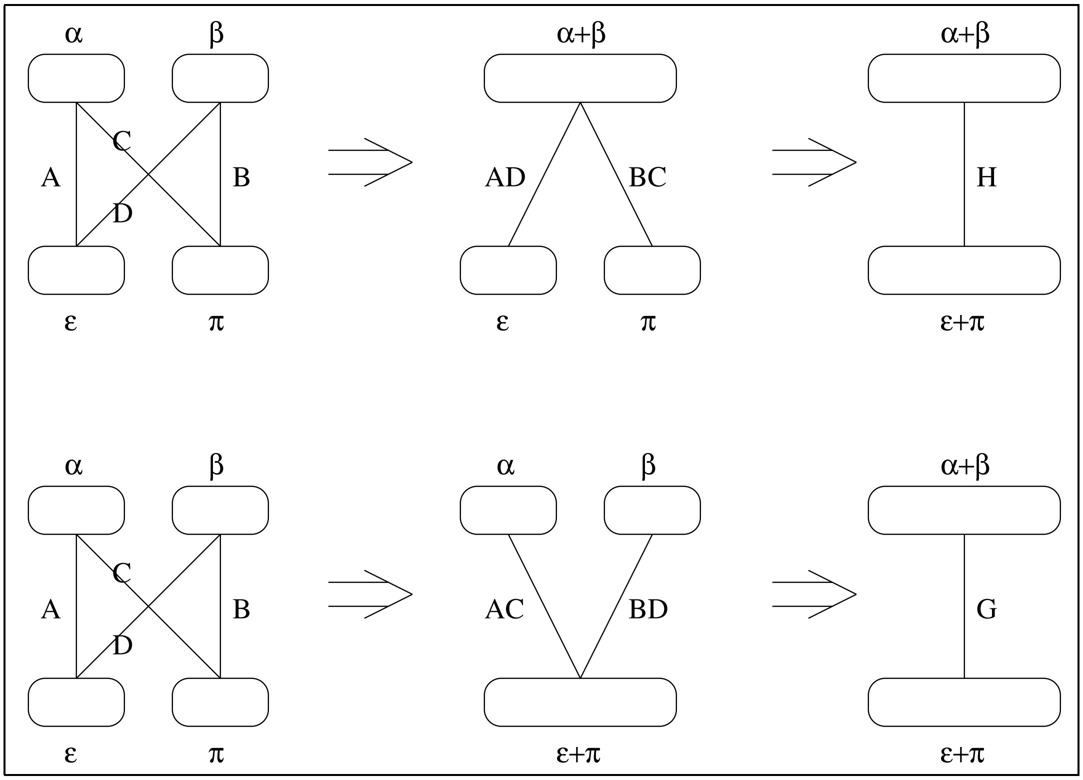


Figure A.5: Symmetric property

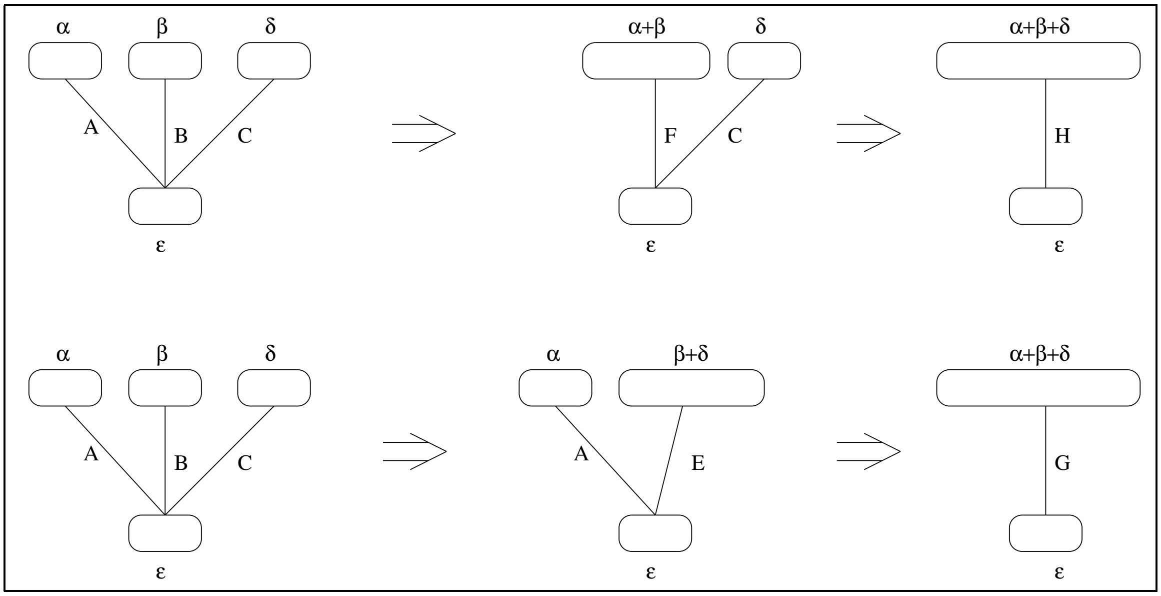


Figure A.6: Wedge associative property

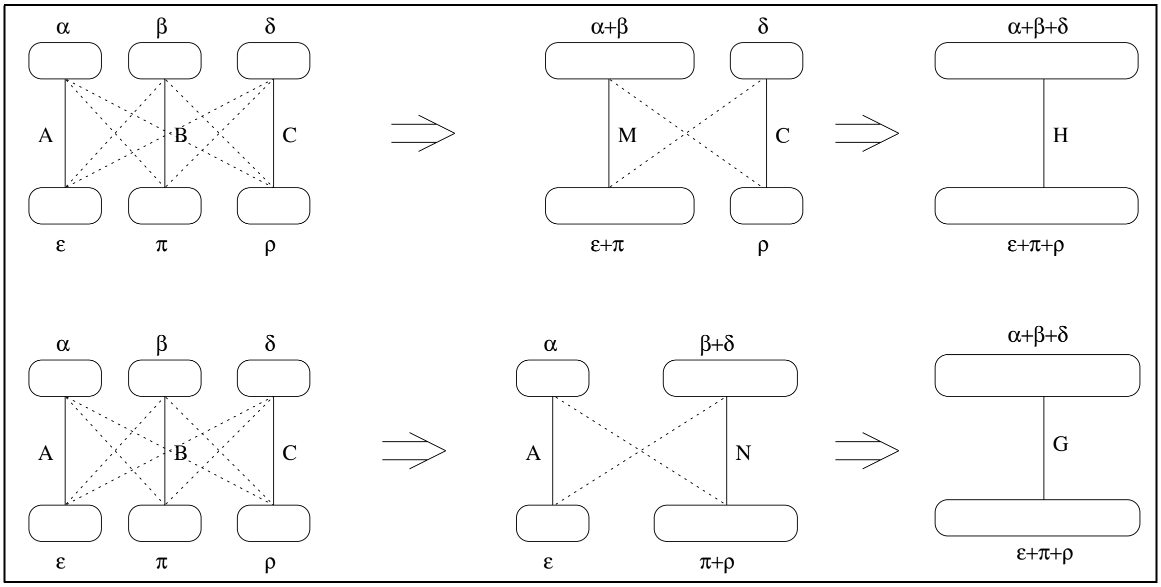


Figure A.7: Associative property

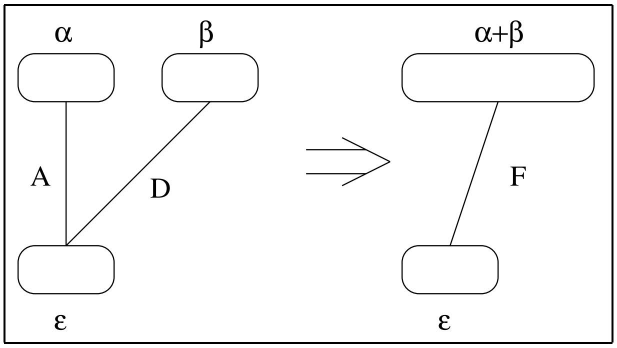


Figure A.8: Merging wedge-type link pair

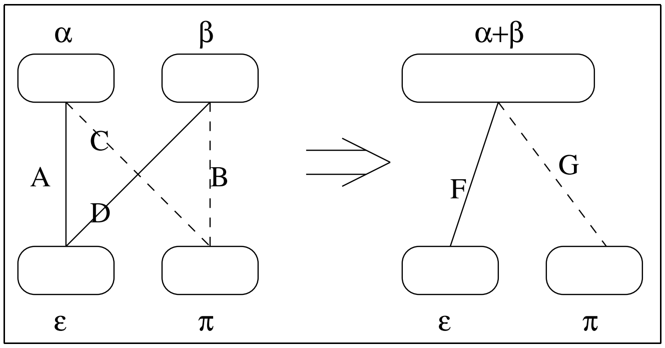


Figure A.9: Merging non-wedge link pairs

The formula used to join links in this thesis is based upon the formula for merging a simple wedge-type link pair. Figure A.8 shows such a pair with weights A and D between endpoints of sizes α, β, and ϵ. The similarity F is calculated as:

$$F = A\frac{\alpha}{\alpha + \beta} + D\frac{\beta}{\alpha + \beta} \tag{A.1}$$

Intuitively, the new similarity is a weighted sum of the original similarities: A contributes to F proportionately to the amount of text its endpoint contributes, and B contributes similarly. (The shift from B to D and from H to F is to make the following discussion simpler.)

A.3 Merging formula

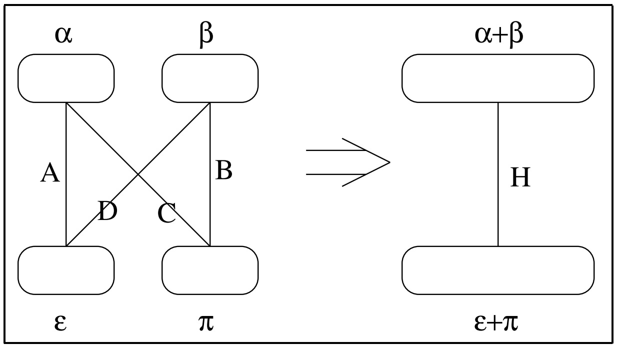


Figure A.10: Standard sample link for proof

The formula for combining non-wedge links is derived from Equation A.1. Consider the four links represented in on the left side of Figure A.9. Using Equation A.1, the links labelled A and D can be merged to create the link labelled F. At the same time, the links labelled B and C must be merged to create the link labelled G. So,

$$\begin{aligned} F &= A\frac{\alpha}{\alpha+\beta} + D\frac{\beta}{\alpha+\beta} \\ G &= C\frac{\alpha}{\alpha+\beta} + B\frac{\beta}{\alpha+\beta} \end{aligned}$$

Equation A.1 is used one more time to merge the links labelled F and G, resulting in the link combining formula yielding a similarity H:

$$\begin{aligned} H &= F\frac{\epsilon}{\epsilon+\pi} + G\frac{\pi}{\epsilon+\pi} \\ &= A\frac{\alpha}{\alpha+\beta}\frac{\epsilon}{\epsilon+\pi} + D\frac{\beta}{\alpha+\beta}\frac{\pi}{\epsilon+\pi} \\ &\qquad +C\frac{\alpha}{\alpha+\beta}\frac{\pi}{\epsilon+\pi} + B\frac{\beta}{\alpha+\beta}\frac{\epsilon}{\epsilon+\pi} \\ &= \frac{A\alpha\epsilon + B\beta\pi + C\alpha\pi + D\beta\epsilon}{(\alpha+\beta)(\epsilon+\pi)} \end{aligned} \qquad \text{(A.2)}$$

The next section demonstrates that these combining formulae provide the properties required.

A.4 Formula satisfies needs

Theorem A.1 *Equations A.1 and A.2 satisfy properties 1 through 7 listed above.*

Proof. By analysis of each property. Unless otherwise noted, the symbols correspond to those in Figure A.10.

1. **Extraneous reduces.** Here $\pi = 0$ since that part is not considered. Also, $B = 0$, $C = 0$, and $D = 0$, the first two because they are not present, the last by definition in this case. With those values, we have from Equation A.1:

$$\begin{aligned} H &= \frac{A\alpha + 0.0 \cdot B}{\alpha + \beta} \\ &= A\frac{\alpha}{\alpha + \beta} \end{aligned}$$

If $\beta = 0$, then $H = A$ and if $\beta > 0$, then $H < A$. (α, β, and ϵ are all lengths and cannot be negative.) Finally, as $\beta \rightarrow \infty$, the denominator dominates the numerator, and $H \rightarrow 0$.

2. **Identical improves.** In this case, $\pi = 0$ again, and $B = 0$, $C = 0$, and $D = 1$. We assume that $A \neq 0$, or this case reduces to a symmetrical version of the previous case. So:

$$\begin{aligned} H &= \frac{A\alpha + 1.0 \cdot \beta}{\alpha + \beta} \\ &= \frac{A\alpha + \beta}{\alpha + \beta} \\ &= A\frac{\alpha + \frac{\beta}{A}}{\alpha + \beta} \end{aligned}$$

But since $0 < A \leq 1$, we know that $\beta/A \geq \beta$, and therefore $\frac{\alpha+\frac{\beta}{A}}{\alpha+\beta} \geq 1$, so $H \geq A$. If $A < 1$, then $\frac{\alpha+\frac{\beta}{A}}{\alpha+\beta} > 1$, so $H > A$. Further,

$$\lim_{\beta \rightarrow \infty} \frac{A\alpha + \beta}{\alpha + \beta} = 1$$

Finally, if $\beta = 0$, then $H = A$.

3. **Partial idempotence.** In this case, consider Figure A.4. Then Equation A.1 yields:

$$\begin{aligned} H &= \frac{A\alpha + A\alpha}{\alpha + \alpha} \\ &= A \end{aligned}$$

which is the desired result.

4. **Commutative** Equation A.1 is commutative by inspection.

5. **Symmetry** Consider the links depicted in Figure A.5. Applying Equation A.1 yields values:

$$\begin{aligned} AD &= \frac{A\alpha + D\beta}{\alpha + \beta} \\ BC &= \frac{C\alpha + B\beta}{\alpha + \beta} \\ AC &= \frac{A\epsilon + C\pi}{\epsilon + \pi} \\ BD &= \frac{D\epsilon + B\pi}{\epsilon + \pi} \end{aligned}$$

   The equation can then be applied to those links to generate:

$$\begin{aligned} H &= \frac{AD\epsilon + BC\pi}{\epsilon + \pi} \\ G &= \frac{AC\alpha + BD\beta}{\alpha + \beta} \\ &= \frac{\frac{A\epsilon+C\pi}{\epsilon+\pi}\alpha + \frac{D\epsilon+B\pi}{\epsilon+\pi}\beta}{\alpha + \beta} \\ &= \frac{A\epsilon\alpha + C\pi\alpha + D\epsilon\beta + B\pi\beta}{(\epsilon + \pi)(\alpha + \beta)} \\ &= \frac{(A\alpha + D\beta)\epsilon + (C\alpha + B\beta)\pi}{(\epsilon + \pi)(\alpha + \beta)} \\ &= \frac{\frac{A\alpha+D\beta}{\alpha+\beta}\epsilon + \frac{C\alpha+B\beta}{\alpha+\beta}\pi}{\epsilon + \pi} \\ &= \frac{AD\epsilon + BC\pi}{\epsilon + \pi} \\ &= H \end{aligned}$$

6. **Wedge associative** From Figure A.6, we get:

$$\begin{aligned} F &= \frac{A\alpha + B\beta}{\alpha + \beta} \\ E &= \frac{B\beta + C\delta}{\beta + \delta} \end{aligned}$$

$$\begin{aligned}
H &= \frac{F(\alpha+\beta)+C\delta}{(\alpha+\beta)+\delta} \\
G &= \frac{A\alpha+E(\beta+\delta)}{\alpha+(\beta+\delta)} \\
&= \frac{A\alpha+B\beta+C\delta}{\alpha+\beta+\delta} \\
&= \frac{\frac{A\alpha+B\beta}{\alpha+\beta}\cdot(\alpha+\beta)+C\delta}{\alpha+\beta+\delta} \\
&= \frac{F(\alpha+\beta)+C\delta}{(\alpha+\beta)+\delta} \\
&= H
\end{aligned}$$

7. **Associative** This property is implied by the *commutative*, *symmetry*, and *wedge associative* properties.

□

Bibliography

[AMY88] Robert M. Akscyn, Donald L. McCracken, and Elise A. Yoder. KMS: A distributed hypermedia system for managing knowledge in organizations. *Communications of the ACM*, 31(7):820–835, 1988.

[Ber90] Mark Bernstein. An apprentice that discovers hypertext links. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 212–223, INRIA, France, 1990.

[Ber94] Berkeley Computer Science Department. The UC Berkeley Technical Report Server. http://cs-tr.cs.berkeley.edu/, 1994. December 15, 1994.

[BSAS94] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART : TREC3. In *The Third Text REtrieval Conference (TREC-3)*, 1994. Forthcoming.

[Bus45] Vannevar Bush. As we may think. *Atlantic Monthly*, pages 101–108, July 1945.

[Cal94] James P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, Dublin City University, Dublin, Ireland, 1994.

[CC92] James P. Callan and W. Bruce Croft. An approach to incorporating CBR concepts in IR systems. In *AAAI Spring Symposium: Case-based reasoning and information retrieval—exploring the opportunities for technology sharing*, 1992.

[CFG91] Michael L. Creech, Dennis F. Freeze, and Martin L. Griss. Using hypertext in selecting reusable software components. In *Hypertext '91 Proceedings*, pages 25–38, San Antonio, Texas, 1991.

[CGN+69] Steven Carmody, Walter Gross, Theodor H. Nelson, David Rice, and Andries van Dam. *A Hypertext editing system for the /360*, pages 291–330. University of Illinois Press, 1969.

[CKP93] Douglass R. Cutting, David R. Karger, and Jan O. Pedersen. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 126–134, Pittsburgh, 1993.

[Con87] Jeff Conklin. Hypertext: An introduction and survey. *IEEE Computing*, 20(9):17–41, 1987.

[Cor94] Cornell Computer Science Department. Cornell University Department of Computer Science. http://www.cs.cornell.edu/, 1994. December 15, 1994.

[CU94] Jack G. Conrad and Mary Hunter Utt. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 260–270, Dublin City University, Dublin, Ireland, 1994.

[Dav93] Clive Davidson. What your database hides away. *New Scientist*, pages 28–31, January 3 1993.

[EE68] Douglas C. Engelbart and William K. English. A research center for augmenting human intellect. In *AFIPS Conference Proceedings: 1968 Fall Joint Computer Conference*, volume 33, pages 395–410, 1968.

[FD92] Peter W. Foltz and Susan T. Dumais. Personalized information delivery: an analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.

[Fou94] National Science Foundation. Nsfnet backbone traffic distribution by service. ftp://nic.merit.edu/nsfnet/statistics/, 1994. November 20, 6:00pm EST. Files history.bytes, 1994/nsf_9405.ports, and 1994/nsf_9410.ports.

[Fri88] Mark E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(7):880–886, 1988.

[FW79] Funk and Wagnalls, editors. *Funk and Wagnalls New Encyclopedia.* New York, 1979. 29 volumes.

[Glu89] Robert J. Glushko. Design issues for multi-document hypertexts. In *Hypertext '89 Proceedings*, pages 51–60, Pittsburgh, 1989.

[Gra94] Matthew Gray. Wow, it's big. http://www.mit.edu:8001/people/-mkgray/wow-its-big.html, 1994. November 20, 1994.

[Hal88] Frank G. Halasz. Reflections on Notecards: seven issues for the next generation of hypermedia systems. *Communications of the ACM*, 31(7):836–852, 1988.

[Har92a] Donna Harman. The DARPA TIPSTER project. *SIGIR Forum*, 26(2):26–28, Fall 1992.

[Har92b] Donna K. Harman, editor. *The first Text REtrieval Conference (TREC-1)*, 1992.

[IB90] W. J. Irler and G. Barbieri. Non-intrusive hypertext anchors and individual colour markings. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 261–273, INRIA, France, 1990.

[Kah91] Brewster Kahle. An information system for corporate users: Wide Area Information Servers. File `wais-corporate-paper.text` from anonymous ftp at think.com, 1991. December 4, 1994.

[Lon76] R. E. Longacre. *An Anatomy of Speech Notions.* The Peter de Ridder Press, Belgium, 1976.

[Lon83] Robert E. Longacre. *The Grammar of Discourse.* Plenum Press, New York, 1983.

[Luc90] Dario Lucarella. A model for hypertext-based information retrieval. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 81–94, INRIA, France, 1990.

[Mau94] Michael Mauldin. Lycos: News. http://fuzine.mt.cs.cmu.edu/mlm/lycos-news.html, 1994. November 20, 1994.

[MI93] Catherine C. Marshall and Frank M. Shipman III. Searching for the missing link: discovering implicit structure in spatial hypertext. In *Hypertext '93 Proceedings*, pages 217–230, Seattle, Washington, 1993.

[MS94] Elke Mittendorf and Peter Schäuble. Document and passage retrieval based on hidden Markov models. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 318–327, Dublin City University, Dublin, Ireland, 1994.

[Nel74] Theodor H. Nelson. *Dream Machines/Computer Lib.* Nelson: Chicago, Chicago, 1974.

[NL94] National Performance Review and Lawrence Livermore National Laboratory. Your ToolKit to help reinvent government. http://www.npr.gov/, 1994. December 15, 1994.

[NLBH88] Debbie Nunn, John Leggett, Craig Boyle, and David Hicks. The rexx project: A case study of automatic hypertext construction. Technical Report 88-021, Hypertext Research Lab, Texas A&M University, April 1988.

[Noi93] Emanuel G. Noik. Exploring large hyperdocuments: Fisheye views of nested networks. In *Hypertext '93 Proceedings*, pages 192–205, Seattle, Washington, 1993.

[Par89] H. Van Dyke Parunak. Hypermedia topologies and user navigation. In *Hypertext '89 Proceedings*, pages 43–50, Pittsburgh, 1989.

[Par91] H. Van Dyke Parunak. Ordering the information graph. In Emily Berk and Joseph Devlin, editors, *Hypertext / Hypermedia Handbook*, chapter 20, pages 299–325. Intertext Publications, McGraw-Hill Publishing Company, 1991.

[PT90] Xavier Pintado and Dennis Tsichritzis. SaTellite: Hypermedia navigation by affinity. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 274–287, INRIA, France, 1990.

[Rei94] Brian Reid. Analysis of stored news articles. Usenet network news list `news.lists` article number 39dufr$oeq$1@usenet.pa.dec.com, 1994. November 4, 1994.

[RS94] Daniela Rus and Kristen Summers. Using white space for automated document structuring. Unpublished, July 1994.

[SA93] Gerard Salton and James Allan. Selective text utilization and text traversal. In *Hypertext '93 Proceedings*, pages 131–144, Seattle, Washington, 1993.

[SAB93] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, Pittsburgh, 1993.

[SABS94] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–1426, June 1994.

[Sal68] Gerard Salton. *Automatic information organization and retrieval.* McGraw-Hill, 1968.

[Sal75] Gerard Salton. *A theory of indexing*, volume 18 of *Regional conference series in applied mathematics.* Society for Industrial and Applied Mathematics, 1975.

[Sal89] Gerard Salton. *Automatic text processing.* Addison-Wesley, 1989.

[Sal90] Gitta B. Salomon. Designing casual-use hypertext: the CHI '89 InfoBooth. In *Empowering People: CHI '90 conference proceedings*, pages 451–458, Seattle, 1990.

[Sal91] Gerard Salton. Developments in automatic text retrieval. *Science*, 253:974–980, August 1991.

[San94] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151, Dublin City University, Dublin, Ireland, 1994.

[SF90] P. David Stotts and Richard Furuta. Hierarchy, composition, scripting languages, and translators for structured hypertext. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 180–193, INRIA, France, 1990.

[SK93] Yoshitaka Shibata and Michiaki Katsumoto. Dynamic hypertext and knowledge agent systems for multimedia information networks. In *Hypertext '93 Proceedings*, pages 83–93, Seattle, Washington, 1993.

[SS94] Gerard Salton and Amit Singhal. Automatic text theme generation and the analysis of text structure. Technical Report TR94-1438, Cornell University, Department of Computer Science, July 1994.

[Stu85] John Stubbs. The new oxford english dictionary and its potential users: Some preliminary comments. In *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 78–81, June 1985.

[The94] The Asylum. The Asylum's staff. http://www.galcit.caltech.edu/t̃a/asylstaff.html, 1994. December 15, 1994.

[Tri83] Randall H. Trigg. *A network-based approach to text handling for the online scientific community.* Ph.D. dissertation, University of Maryland, 1983. Also technical report TR-1346.

[TW87] Randall H. Trigg and Mark Weiser. TEXTNET: A network-based approach to text handling. *acm Transactions on Office Information Systems*, 4(1):1–23, January 1987.

[vD88] Andries van Dam. Hypertext '87 keynote address. *Communications of the ACM*, 31(7):887–895, 1988.

[Voo93] Ellen M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180, Pittsburgh, 1993.

[Wil90] Eve Wilson. Links and structures in hypertext databases for law. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 194–211, INRIA, France, 1990.

[Wil94] Ross Wilkinson. Effective retrieval of structured documents. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317, Dublin City University, Dublin, Ireland, 1994.

[Wit53] Ludwig Wittgenstein. *Philosophical investigations.* Basil Blackwell and Mott, Ltd, Oxford, England, 1953.

[XBC94] Jinxi Xu, John Broglio, and Bruce Croft. The design and implementation of a part of speech tagger for English. Technical Report TR94-26, University of Massachusetts, Department of Computer Science, 1994.

[YGM93] Tak W. Yan and Hector Garcia-Molina. Index structures for information filtering under the vector space model. Technical Report TR-1494, Stanford University, Department of Computer Science, November 1993.

[You90] Laura De Young. Links considered harmful. In *Hypertext: Concepts, systems and applications: Proceedings of the European conference on Hypertext*, pages 238–249, INRIA, France, 1990.