RELATED PROBLEMS OF KNOWLEDGE DISCOVERY

Judith P. Carlisle  
*Texas A&M University*  
j-carlisle@tamu.edu  

Scott W. Cunningham  
*NCR Corporation*  
scott.cunningham@atlantaga.ncr.com  

Anustup Nayak  
*Georgia Institute of Technology*  
anustup@hotmail.com  

Alan Porter  
*Georgia Institute of Technology*  
alan.porter@isye.gatech.edu  

Abstract

Research within the communities of informatics, scientometrics, cybermetrics, and knowledge discovery has, until recently, occurred in comparative isolation. In spite of this isolation, there are common concerns and techniques that would benefit from a cross-fertilization of ideas. In this paper, these research areas are defined in terms of data, application, and discipline. A shared set of analysis tools and theoretical approaches are identified and presented toward the goal of establishing a shared theoretical basis for the disciplinary problems.

1. Introduction

Bibliometrics, scientometrics, informatics, and cybermetrics are each highly developed fields with distinct and proprietary methods for the handling of content and the treatment of documents. However, an analysis of their respective techniques reveals that they are more similar to one another than previously understood. In this paper, such an analysis is presented, resulting in an improved understanding of the techniques and principles which are common to these fields. In this paper, we present this result as a common framework which can be useful when applied to the problems encountered in KDD. This framework can contribute to the establishment of a firm theoretical basis for KDD in text databases.

The paper is organized as follows. In the first section the disciplinary backgrounds of target fields for the analysis are discussed. The second section of the paper focuses on the meaning of content and document as applied to these disciplines. Section three explores the nature of data found in each of the disciplines. In section four, we present an analysis emphasizing the identification of a common solution. A summary is provided in section five.

1.1 Disciplinary backgrounds

In this section, each of the functional areas is identified and defined as it pertains to KDD related tasks. These functional areas are: informatics, scientometrics, bibliometrics, and cybermetrics. Toward the goal of achieving a correlation between techniques and applications, we focus most closely on the common disciplinary research on documents and document classifications. Although we broadly characterize the research in these areas, it is certainly the case that there is a considerable amount of interdisciplinary research and development.

Informatics, AKA information science, is the study of information, especially the collection, management, and retrieval of informative texts. Informatics historically relies on quantitative content analysis techniques of language, correlating the similarity of documents to similarity of word occurrence statistics and co-occurrence patterns. Informatics is useful to applications such as information retrieval (IR), where quantitative representations of text provide rich and concise insight into the information content of documents. Informatics is most often applied in the analysis of content within documents and allows for effective comparison of information contained within the documents. Pioneers in informatics have discussed the design of effective information retrieval systems using such informatic measures [10]; [11].

"Cybermetrics" [3] is a newly coined term for the discipline of analyzing Internet-related data sources. Also sharing research interest in large, electronic and interactive data sources is the discipline of knowledge discovery in databases [7]. This discipline combines
techniques from machine learning, artificial intelligence and statistics to uncover patterns in large databases. While concerned in large part with numerical data sets, text databases remain an important and visible application area for knowledge discovery.

Scientometrics is the quantitative study of the processes and outputs of scientific activity. Although not exclusively concerned with scientific content, understanding the patterns of scientific activity between and within disciplines is an important component of scientometric research. De Sola Price is widely held to be the founder of the modern discipline of scientometrics [9]. Scientometrics, then, emphasizes measures of relatedness of content --- the document is at best an artifact or partial indicator of research activity.

Bibliometrics is the quantitative study of relationships between documents within a document collection. While bibliometric data includes identification and analysis of document content, bibliometrics also includes non-content-related data such as author, source, co-authoring activities, etc. Pioneering work was done [1] which formulated a law describing the distribution of content across documents. Bibliometric techniques are particularly valuable for use in library science and collections management. The applicability of bibliometric indicators has expanded from the analysis of electronic indices of paper-and-print documents to the analysis of fully electronic document collections such as the Internet. Though KDD differs with regard to the type of problems to which it can be applied, bibliometrics and KDD techniques are closely related.

In order to explore an underlying theoretical basis of KDD, the rest of the paper is organized as follows. To establish a perspective to view the relationship between the three disciplines, Section establishes a "content to document" framework. Section 3 utilizes this framework to examine applications which are characterized as intradocument, interdocument, and intracorpus. Section 4 examines problem solutions, which the three disciplinary areas have in common. Finally, Section 5 recapitulates and identifies future research directions to which this paper may be applied.

2. Content, document, and KDD

The three research disciplines identified are all similar in that they focus on the extraction, representation, and analysis of information which is textually represented. While each of these disciplines have similar data source, they each have a different focus of the data analysis intent. These differences can be described using the ideas of content and document, where content refers to the information, which is textually represented, and document refers to characteristics of a document, which contains information.

Table 1 depicts each of these fields, along with its position along a content/document grid and an example application. Apply the following legend: C-by-D should be read as "content-by-document," C-by-C as "content-by-content," and D-by-D as "document-by-document.

<table>
<thead>
<tr>
<th>Data</th>
<th>Application</th>
<th>Discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-by-D</td>
<td>Information Retrieval</td>
<td>Informatics: Cybermetrics</td>
</tr>
<tr>
<td></td>
<td>Internet Search Engines</td>
<td></td>
</tr>
<tr>
<td>C-by-C</td>
<td>Science Maps: AutomatedSciometrics Classification</td>
<td></td>
</tr>
<tr>
<td>D-by-D</td>
<td>Query-by-Example</td>
<td>Bibliometrics: KDD</td>
</tr>
</tbody>
</table>

For example, information retrieval is a content by document discipline --- both content and content origin are important to the application. The term "by" is used to denote the relationship between the two dimensions of the data matrix. Content of documents is used to find sources of content. Using various techniques, such as inverted file indexing, Bayesian probability algorithms, or cosine alignment, documents within a document corpus which contain prespecified content may be identified. Content by content data is used by scientometrics to identify trends, patterns, or other analysis of content contained within a document. Such applications emphasize information content within a document collection rather than the source of content. In other words, content is analyzed outside of the context of the document in which it is contained. Document by document disciplines, such as bibliometrics --- and KDD --- emphasize analysis of source of document, that is documents, in order to ascertain hidden, useful relationships between documents. In order to
elucidate, the notions of content and document are more full explicated.

2.1 The meaning of content

Fundamental to each of these disciplines is the notion of content. Content refers to the kind or quality of information imparted by specific documents or collections of documents. The challenge for information retrieval system design is to produce systems which consistently retrieve and return information deemed relevant by specific communities of users. Content, therefore, is a relative term, judged only by prospective human raters.

Written language is one means of imparting content; select words and phrases can be serviceable indicators of document content. One challenge for IR system development and use if the reconstruction of contextual information in a surrogate document representation which omits much semantic information. Another challenge to effective IR is the indicators of document content. One challenge for IR content: select words and phrases can be serviceable for word meanings and semantics.

2.2 The meaning of document

Equally important to the three disciplines is the notion of a document. Each discipline uses the documents in varied fashions. A document maybe exclusively paper-and-print, an electronic reference to a paper-and-print document, or a fully electronic document. A document may vary in the extent of its content: it maybe an abstract of a longer work, it may be a paper entire, or it might be a book. A document might also vary by media: some might by exclusively text, others hypertext, still others multimedia. However, ultimately document refers to the overriding prevalence of information presented using language represented via alphanumeric text.

Regardless of the form of the document, it is necessary to establish an index so that content (and whatever words or phrases signify content) is indexed across the document corpus. Regardless of the format of the original source document, this indexing is inevitably done electronically. The size of the collection (and growing complexity of retrieval operations) requires the use of computers for effective analysis and retrieval. The varied media of the original document is of concern only in that written text may be only a partial indicator off the actual document content. The final issue is the extent of content seen across documents --- some documents are inevitably more richly indexed than others. Analysis techniques, regardless of their area of application or theoretical origins, must deal with the heterogeneity of indices.

The fundamental data type, used across disciplines, involves an index of content across documents. This representation can be relatively straight forward, such as a matrix of occurrences, or a more complex system such as an index schema for sparse matrices. In following section, use of this fundamental data type across seemingly very different applications is considered.

3. The nature of the data

In this section, the nature of data is examined. Three function of information analysis and application are identified: interdocument, intradocument, and intracorpus. Intradocument utilizes language within a single document to establish significant content of that document. Interdocument, which is referred to as language activity, examines relationships and activity between language-presented information. Interdocument disciplines results in knowledge about language patterns such the relationships between words and activities (such as research) which make use of those words. This type of knowledge allows similarities about and trends in formation content of a document collection to emerge. Intracorpus, language rules, allows for significant patterns across an entire document collection to emerge for analysis and mining activities. Intracorpus differs from interdocument in that information other than content is recognized as relevant and useful to analysis. Examples of such information include, but are not

---

1 Salton's term frequency/inverse document frequency measure to eliminate noise from a surrogate document representation space [10].
limited to, co-authorship, area of research disciplines, time and place of publication, and institution involvement in content area. Intradocument application applications are useful for intradocument analysis, interdocument for interdocument analysis, and intracorpus for intracorpus analysis.

**TABLE 2**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Knowledge Discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intradocument</td>
<td>Content/Language Informatics</td>
</tr>
<tr>
<td>Interdocument</td>
<td>Relationships/ Activity Scientrometrics, Bibliometrics, Cybermetrics</td>
</tr>
<tr>
<td>Intracorpus</td>
<td>Patterns/Rules Bibliometrics Knowledge Discovery</td>
</tr>
</tbody>
</table>

Table 2 rates the various disciplines according to the kinds of comparisons typically made (intradocument, interdocument, intracorpus), and the kinds of knowledge being sought from the comparison. We note that informatics is typically (though not exclusively) focused on language and content within individual documents. In contrast, scientometrics and bibliometrics is more concerned with relationships and activities across multiple documents. More exhaustive again is KDD which seeks broad rules and patterns. KDD also values information external to the corpus such as business knowledge or practices. We now turn to the basic data types used --- these data types are necessary regardless of the specific discipline or application. Each application area from Table 1 can be viewed from the perspective of whether it is inter/intradocument or intracorpus.

### 3.1 The information retrieval application

Information retrieval is an intradocument and interdocument application. Intradocument because it seeks to represent content of each document within a document collection (including queries) as a dense index of some type (inverted file, vector, enhanced Boolean). An underlying assumption is made that those words which occur with significant frequency within a document are the "important words," i.e. those words which best convey a document's content. Because high frequency terms are often the most noisy, these terms (as identified through techniques such as Salton's tf/idf) are removed from the content representation.

### 3.2 The science mapping application

Science maps are two-dimensional (or higher) representations of bibliometric content objects, their relative importance, and their linkage. Cartographic representation of scientific information alludes to the relative important and degree of similarity (relatedness, or connectivity) between the content objects. For example, the map may represent the relative emphasis of subtopic areas in a particular discipline. The positioning of the subject areas and their connections to each other could indicate shared concepts or transfer of ideas. Science maps can be drawn by quantitative analysis followed by visualization of bibliographic content.

### 3.3 The query-by-example and case-based reasoning applications

Both query-by-example and case-based reasoning are examples of intracorpus, or intracorpus, applications. Query-by-example utilizes an example of desired information, such as a document, and attempts to find other documents within a corpus which have similar content as well as relationships between significant components of content. Hence, content is analyzed in the context of the document. Similarly, case-based reasoning views the document as an exemplar, and seeks to identify similar cases within a document corpus. These methods rely on measures of similarity, relatedness, or distance to make retrieval possible.

### 4. Common solutions

The retrieval space in information retrieval can be described thus: a collection of documents (D) consist of a set of documents (d₁, ..., dₖ) where j represents the number of documents in the collection. Content of the document space is represented by a set of words, or word token, (W) such that (w₁, ..., wₖ) where j is the number of all words contained in all documents. Then, in a simple Boolean matrix, for example, a document space can be represented by a two-dimensional matrix with D along one axis and W along the other. Presence of a word in a document, or absence of a word from a document, is respectively represented by a 1 or 0. Figure 1 shows such a document space represented as the matrix X.
Figure 1. Basic data - content indexed by document

The matrix X depicted in Figure 1 has the dimension of document-to-content. We may use an index I into the documents, and an index J into the content. A particular row of the matrix X(i,J) displays all of the content associated with a given document (i). A particular column of the matrix X(i,J) shows which documents contain a specific item of content (j). Further data types can be derived from this matrix by multiplication. A matrix on content-by-content can be obtained by post-multiplying the matrix by its transpose. For this figure, read "content" along the axes marked by "C" and "document" for those indicated by "D".

Figure 2. Derived data --- content-by-content, document-by-document

Figure 2 illustrates this using the original matrix X and its transpose X\. This symmetric matrix has dimensions of content-by-content, and has the same index j for both rows and columns. This matrix may be used to determine the co-occurrences of particular items of content across the document collection. This symmetric matrix can be used to determine the number of shared items of content held between two documents as shown below in Figure 3.

Figure 3. Higher-order relationships

It is possible to derive other higher-order or indirect relationships from this matrix, as depicted in Figure 3. For instance, the number of links representing the existence of shared content between two documents can be determined. Such links may be either direct or direct reflecting the existence of reflexive, transitive, and/or symmetrical relationships with the content as illustrated in figure 3 where three documents share content: topics a, b, and c. Documents one and two are linked directly (via topic a), as well as indirectly (via topics shared with document 2).

The document-by-document matrix can be derived from the multiplication of the matrix transpose (X^T) with the matrix X (see Figure 2). Such multiplications can be continued. The higher order matrix (X^T*X^T*X) shows all the linkages possible between documents when one intermediate step is allowed through other, related documents. This computation of linkages can continue indefinitely through successively high-order matrix multiplications. [11] discusses the use and interpretation of these higher-order matrices in information retrieval. In his 1968 work, Salton proposes various transformations to these basic data types [10].

We have now seen three data types: content-by-document, content-by-content, and document-by-document. Each of these data types is idempotent --- meaning they are invariant to multiplication. This means that the basic content of the matrices remains unaltered regardless of how many multiplications of the data are applied [5]. Because of the underlying relationships between the matrices, the disparate approaches in each of the three applications can be
unified. Theoretical comparison between disciplines are possible and valid.

4.1 Matrix structure and data reduction

We have seen, then, how the content-document matrix is a fundamental data type from which other types of data can be derived. This data, once derived, is invariably reduced using one of a variety of statistical or machine-learning techniques. The rationale varies across application areas. Users of content-by-document data seek hidden structures among synonymous words in the data [5]. Users of content-by-content data typically require a low dimensions, easily visualized view of the data (Healey et al. 1996). Document-by-document data collection is often exhaustively large, and requires data reduction for easy storage and analysis. It is also often the case that the content or features of the data are intercorrelated, and an unbiased comparison across the features is necessary table 3 lists these techniques.

Table 3. Data types and their reduction

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Rationale for Dimensionality Reduction</th>
<th>Favored Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content by Document</td>
<td>Synonymous words hidden in structure</td>
<td>Latent Semantic Indexing, Principle Components Analysis</td>
</tr>
<tr>
<td>Content by Content</td>
<td>Low dimensional view onto the data</td>
<td>Multidimensional Scaling</td>
</tr>
<tr>
<td>Document by Document</td>
<td>Unbiased comparison among features; data reduction</td>
<td>(Various)</td>
</tr>
</tbody>
</table>

Figure 4 is used to emphasize the rationale for chosen reduction techniques according to discipline, or application area where U represents the reduced document space and V represents the reduced semantic data space. In this section, we discuss effects of reduction techniques. The favored techniques for data reduction vary. To illustrate, we consistently utilize the single value decomposition (SVD) techniques [8] --- a linear approach to data reduction.

For content-by-document applications, discover of similar word profiles is a key element. Synonymous and similar words are "put closer together" in the reduced semantic space. Hence, identification of conceptual similarities are more readily apparent for identification purposes.

Content-by-content applications, such as science mapping [12], emphasize the identification of emerging trends, patterns, and other similar data within a document corpus. Hence, the need to create an easily manipulated semantic space is emphasized and achieved via matrix multiplication. The resulting information is sufficient for applications such as policy and internal criteria of science, for example, a question such as "how does funding of one area of science affect another, perhaps unrelated, branch of science," (Weinberg). However, content-by-content applications do not limit analysis to presence of content only. The content source (individual documents) can be identified for source verification. For example, an insight not yet recognized by the scientometric community involves the ability to move from document (U) analysis to semantic (V) analysis [4].

Finally, we have the document-by-document reliant applications. These applications areas are typified by analysis such as case-based reasoning, where a particular document can be viewed as a case instance. Hence, correlation for document content requires full content depiction augmented by robust depiction of
logical relationships between content markers. Document-by-document applications require document features, both dependent upon and independent of, one another to be easily assessed so that sophisticated analysis and assessment of full document features within the context of the unique document can be undertaken. Another example of a document-by-document type of application is query-by-example, where query results can be used as examples of other desired information.

5. Conclusion

In this paper we have presented the rudiments of an analysis framework for KDD focusing on its relationship to other language-represented information assessment disciplines. Our goal for the paper is to show diverse communities that significant work relevant to KDD has been done and to provide common conceptual structures based on concepts which already exist in fields that participate in KDD research and development. Toward this end, pertinent data is described via a content/document framework and applications identified within the inter/intradocument and intracorpus framework utilizing their data source as the basis for this division.

KDD is, by nature, interdisciplinary and attracts broad and diverse research and user communities. The content/document framework provides a common conceptual ground to facilitate the ability to share and reuse concepts and results across participants.

We have found this framework to be useful to coordinating the efforts in complex domains such as technology innovation forecasting, a research area which requires the collaboration of a diverse community of research and the development of sophisticated data analysis techniques over a broad spectrum of distributed repositories of network information. Other examples of such complex topics include digital library initiatives, electronic commerce, and virtual communities.

References


