Report on DIMACS Working Group on Data Analysis and Digital Libraries

Meeting of May 17, 2001

Author: Paul B. Kantor (Rutgers University)
On behalf of the Steering Committee: Elizabeth Liddy (Syracuse University), Warran Greiff (Mitre Corp.), Michael Berry (Univ. of Tennessee), Endre Boros (Rutgers Univ.), and David Hull (WhizBang)

A. Workshop Focus

The goal of the DIMACS Working Group on Data Mining and Digital Libraries was to bring together researchers whose work bears on the following question: "Since the content of a Digital Library is machine readable, to what extent can the computers that maintain the library also extract new and useful information from that content?"

The workshop aimed at better communication among members of the diverse communities which all have something important to offer about the problem. A key challenge in the course of the meeting was to develop a common language, enabling workshop participants to bring their own knowledge to bear on the problems that interest others.

B. WORKSHOP TOPICS

Among the application areas explored were:

- clustering of documents and collections
- collection selection
- topic detection
- adaptive filtering
- automatic thesaurus construction

The tools discussed included:

- vector models and linear classifiers
- Boolean models and rule based models
- probabilistic models of relevance
- inference networks for information retrieval
- human rule finding capabilities
- bigram and n-gram based models for text
- NLP based models for organizing and mining libraries
- co-citation and other bibliographic linkages
- web-based and link-based methods of organization

This workshop welcomed researchers who are addressing one or more of these problems: image indexing, sound indexing, text indexing, clustering (of image, sound, text); collection selection; information retrieval; pattern finding in text collections. We were also interested in new research on the problem of evaluating schemes for either retrieval or clustering and classification that can bring some much needed rigor to this problem.

We also welcomed discussion of the issue of "evaluating" systems. Two practically oriented "gold standards" have emerged. One is the TREC setting, an ARPA/NIST/ARDA sponsored annual workshop in which dozens of systems address the same fixed set of tasks, with performance on those tasks being assessed impartially at NIST. The other is the commercial web setting, where schemes compete to attract the eyeballs of surfers, and the methods are often tailored to hold the user's attention, while providing acceptable levels of organization and retrieval. The effectiveness (in, say, the TREC senses) of the commercial schemes is not well understood. Perhaps neither set of measures will ultimately prove most effective in defining the new technology of digital libraries.

Presenters were encouraged to present both recent results and open questions. In keeping with DIMACS tradition, we were most interested in mathematically rigorous approaches to these problems. We recognized that practice may precede theory, and we also welcomed papers presenting methods which prove very effective, even though the theoretical justification is unsatisfactory.

C. TALK ABSTRACTS AND SUMMARIES

The following give some speaker-prepared abstracts, mildly edited, and our own summaries of the presentations.

Statistical Modeling for The Segmentation of Archived News Broadcasts, Warren Greiff, MITRE

Abstract:

Not all information streams come with individual items delimited. For example, broadcast news captured from the air waves may not be accompanied by indications of where one story ends and another begins. In this talk we present recent work carried out at The MITRE Corporation on the automatic segmentation of textual data. The approach taken is based on a fine-grained Hidden Markov Model. Because of the very large number of parameters in the model, robust parameter estimation is a major concern. Since the model's parameter values can be expected to vary smoothly as a function of time, we have based the determination of conditional probabilities on non-parametric density estimation. In this presentation, we will discuss the design of the Hidden Markov Model; reduction of the input stream to the feature vectors which are treated as observables in the model; and the non-parametric estimation techniques we employed.

Summary Notes on Workshop Presentation*:*

In order to capture the broadcast news, the broadcast stream is sent first to an automatic speech recognition module and then (or directly) to the story segmentation module. After segmentation, the output is added to a database, which can be finally accessed by the user.

The adaptive approach to segmentation can be briefly described as follows: The unsegmented source sends streams to a segmenter, which provides segmented output. The segmenter is created on a statistical model, based on a training model obtained from a training set of data.

The MITRE Segmentation System used in 1999 was based on a Naïve Bayes classifiers model, took independent boundary/non-boundary decisions at each word, and had the following features: overlap (the amount of overlap and the overlap weighted by rarity), delta-overlap, pre/post-trigger cues (for words that appear at the end/start of a story), and X-duration cue (i.e. the duration of untransmissible input, e.g., music).

The current MITRE 2000 segmenter is based on a fine-grained Hidden Markov Model (HMM). The coherence features of the current system -- window and spacing -- indicate if a certain word at the present point appeared in the past, usually in a 50-word length buffer. There are 4 types of coherence features: Coher1 (considers the last 50 words in the buffer); Coher2 (considers a buffer of 50 words beginning with 50 words back), Coher3 (considers a buffer of 50 words beginning with 100 words back), Coher4 (considers a buffer of 50 words beginning with 150 words back). This represents a discretization of an autocorrelation relation among terms.

The trigger features of the current MITRE segmentation system look for words appearing more frequently in specified regions (near a boundary).

The initial topology of HMM modeled the middle of the story well, but did not do so well at the start and the end. Consequently, the initial topology created some problems, for example confusion about end of short stories, and reduction of the values of the coherence features. The current version of HMM solves the above problems by disregarding the end-of-story states.

The probability estimation problem consists in estimating the probabilities for each state, and the observed probabilities for each possible transition. The probability estimation problem is solved by considering a smoothed histogram, from which conditioned probabilities $P(S_{t+1}=0|S_t=s)$ are derived. The smoothing approach requires the following assumptions: the transition probabilities don't vary much from state to state, and for a given feature, the observed probabilities should be similar for neighboring values, as well as in neighboring states.

Non-parametric estimation of the probabilities assumes that the values are binned, and produces the smoothed form for every binned feature value f_V, either by performing a non-parametric regression, or by normalizing smoothed combinations of the form $p(f_V|s)$.

The actual segmentation system uses the Viterbi algorithm (maximum likelihood state sequence) and output boundary each time the word generated from the state is 1.

Some directions for future work include smoothing procedures (e.g., 3-dimensional), features (for e.g., directly extracted from Audio), and generalizing the HMM topology.

Collections Of Documents And Questions. A Unified Perspective. (Bicorpa, Features, and Other Animals), Paul Kantor, Rutgers University

Summary Notes on Workshop Presentation*:*

In Data Analysis, there are several methods for the prediction of the CR (credit risk) from sets of known data. For example, in the real case: regression, ANOVA; in the Boolean case: Logical Analysis of Data (LAD), C4.5 rules; support vectors, naïve Bayesian, etc.

For Digital Libraries (considered as collections of books), the easy Gold Standard (1) can be formulated as follows: the human analyst (librarian) perceives the collection as a whole, catalogs books, subcollections, and knows what subcollections are present; the elements of the collection are characterized by multiple attributes, for example genre, subject, authenticity, timeliness.

The hard Gold Standard (2) requires that the human analyst be a subject specialist/scholar, to work in a subject based subcollection, to know if there is something of interest in a certain subcollection, to know the existing theories, and to detect topics, contradictions, summaries, etc. There are no solutions for the Gold Standard (2), just some ideas, which will be presented in the following.

First, consider the question: "How can a collection understand itself"?
Let D be the set of documents, d elements of D; F the set of features (attributes) and f elements of F. We consider the vector representation of d by attributes d -> {w(d,f)}, where D and F are seen as "different entities, and w:DxF -> Reals.

An alternative approach formulates a **direct sum** model, so that matrices decompose as follows.

	F	D
F	Sim(f,f')	W^T(f,d)
D	W(d,f)	Sim(f,f')

In **activation** models, we replace the main diagonal entries by unit matrices:

	F	D
F	1	W^T(f,d)
D	W(d,f)	1

Continuing the activation, if M is the activation matrix

	F	D
F	1	W^T(f,d)
D	W(d,f)	1

then we may compute the powers M, M^2, M^3 etc. These iteration procedures yield relations within either subspace, F or D. Moreover, we sum up the interactions to any order n, with an attenuation factor α and obtain $M^{[n]}$=[(α M)^{n+1}-1]/((α M)-1)

In order to reduce a collection to a subcollection, we may use $M^{[n]}$ to define subsets, by replacing the elements with 0 or 1, according to a certain threshold.

The above approach can also be applied to a set of queries, users, or quests. The ultimate goal is to build a "self-aware" collection, for which the key step is to make a wise choice of F (since several choices are possible). After that, use either method to partition the document collection.

A harder question, which can be raised, is: "Can we learn to recognize contradictions within a collection"? That might be approached with cleverly designed training sets, but it is not clear how many kinds of "contradiction" there are to be discovered.

Data-Mining, MetaData, and Digital Libraries, Elizabeth Liddy, Syracuse University

Abstract:

A Digital Library is a collection of digital objects (the Repository), descriptions of these objects (Metadata), to provide search, browsing, and retrieval (Services), to a distributed set of users (Community). As such, it can be seen that Digital Libraries are an anachronistic phenomenon that attempt to combine the most traditional aspects of

libraries with the most advanced of computer capabilities. This potential will only be adequately fulfilled, however, if digital libraries can take their age-old reliance on human catalogers, who have traditionally provided the searchable descriptions of library materials, and replace them with an automated means for determining and assigning substantive descriptions of digital objects as they are added to a Digital Library. NLP-based Data-Mining software can accomplish the automated metatagging of digital materials by interpreting the textual documents (or their abstracts or introductions) utilizing algorithms which extract meaning at all the levels of language at which human catalogers interpret what a document is about and describe it for effective access by Digital Library users. By utilizing Data-Mining to automatically produce MetaData, there is really no limitation to the number and size of Digital Libraries.]

Summary Notes on Workshop Presentation*:*

By definition, Data Mining (DM) is the computational process of extracting useful information from data, and has as goals to map the low-level data into a richer, more abstract representation, to detect meaningful patterns. It is typically conducted on structured data bases. Data Mining can be extended to Text Mining, by processing the text. The Natural Language Process (NLP) can provide necessary techniques for Text Mining.

Text Mining is the process of analyzing naturally occurring text, in order to extract (tag) vital information. After the analysis, the text is either stored in a database for exploration, or it is automatically (meta)tagged from the documents.

Metadata (MD), originally based on the MARC format, was a record-oriented language with formats for bibliographical entries, and did not support features needed for full-text documents. Metadata moved to SGML, which confers a more flexible format, syntax, philosophy, and the information is tagged meaningful. The ' Dublin core + ' metadata scheme and an educational object serves as an example.

NLP can be used to process data, by performing morphological analysis, lexical analysis, syntactic analysis, or semantic analysis.

Regarding extraction for ML-based classification, NLP should extract Semantic Relations. ML-based Text Classification is performed by any of the following methods: regression, nearest neighborhood classification, Bayesian probabilistic classification, or decision trees.

The Automatic MetaData Generation Methods are the following: Structures Information Extraction (SIE), Textual Information Extraction (TIE), and Text Characterization (TC).

The Digital Library (DL) consists of a collection of digital objects (Repository), with description of these objects (Metadata), which provide search services (Services) to a set of users (Community).

Up to now, we know that NLP is used in DM; DM is used to obtain MD, and MD is used in DL.

The research issues for Era 1 are the following:

- What is the appropriate set of metatags to be assigned to objects in order to become accessible by users?

- How to determine this set?

- It is known that controlled vocabulary is worse than free-text indexing; what retrieval performance can DL expect from metadata and tags?

- Can NLP automatically produce MD for all DL services?

The Research Issues for Era 2 are the following:

- How to extend DM applications to understand both the content and the users of DL?

- What MD needs to be generated?

Using User Data To Help Improve IR Functions In The Context Of Web Search, Mark Hansen, Bell Labs
(Joint work with Elizabeth Shriver, Bell Labs, and Ben Rubin, EAR Studio)

<u>Summary Notes on Workshop Presentation</u>*:*

Some recent projects at Bell Labs can be described as:

- Characterize navigation patterns (i.e., predict the accessed resources, suggest schemes from context, create a user profile based on navigation, etc.)
- Improve web searching, by reducing the number of resources, simplifying the resource lists (automated characterization), boosting likely relevance, etc.
- Monitor and analyze streaming context, by implementing large-scale data collections, assessing flow of information in network communities, providing customizable monitors, developing audio interfaces, etc.

To improve searching, Web pages are characterized by page content and link structure. Special collections can be accessed using subject-specific search engines, and manually derived/maintained directories. The user feedback of relevance requires specialized applications, and opt-in user participation. The server logs: click through statistics; proxy or ISP request logs.

The goal of the "improve web searching project" is to find information faster. To accomplish this goal, the web search technique should borrow strength from auxiliary data (for example, records of the user navigation via proxy or gateway access logs, existing content hierarchies, page structure, user profiles). Regarding the implementation, the search engines should extend the current ones, and use a proxy ISP.

The motivation for query clustering is that across search sessions, similar URL's could be found. The proxy or ISP display augments a search engine with URL's suggested by previous activity. The proxy implementation of the user model describes a searching task, and provides template for mining proxy logs. The search session extraction begins with a query to a search engine. Then it extracts URL's from result pages, and examines links on required pages to identify connected paths (keeping track of all items).

The idea of improving the web search is to create query groups, based on URL's visited during a search session, and to improve the URL ranking. The goal is to associate weights for the queries and URLs, which are accessed by certain groups, and to obtain triplets of the form (qm, group_id, weight), (URL, group_id, weight).

The methodology for search improving is a model-based clustering, which uses a "generative" probability model for search engines. The model components can be briefly described as follows: randomly select a query group; sample URL in the search session, and output (qm, group_id, weight), (URL, group_id, weight).

Parameter estimation is performed using the Expectation Maximization(EM) technique. There are several variant of EM algorithms: The incremental EM algorithm updates only the parameters from URL's belonging to the queries, and it is characterized by a fast convergence. The on-line EM updates only parameters in active sessions, it is not iterative, and performs a dynamic addition/deletion of clusters.

Incorporating labeled data is a solution for eliminating the following limitations of the mixture model: the sensitivity to noise, and the fact that infrequently accessed sites can be swept into large clusters, whether or not they belong there. The hierarchies of URL's consist of manually created directories, and open source topic trees (dmoz).

The validation of the improved model was done on the Lucent proxy access logs, with 2/3 training data and 1/3 testing data. The effectiveness metric, based on the assumption that the desired URL's are in the testing data, outputs the number of pages the user has to investigate and the percent of queries for which the users display the desired URL's.

Summarizing the validation results, the new method which incorporates labeled data reduces effort in searching by producing shorter lists (85% coverage with labeled data, 60% without labeled data), fewer URL's examined (8 versus 4.1 with labeled data).

Monitoring And Triage Of Business News, Foster Provost, New York University
(Joint work with R. Sankaranayanan, New York University, Hirsh, V. Dhar, S. Mackassy, Rutgers University)

<u>Abstract</u>:

We are interested in systems that monitor information feeds and perform triage on the information in order to improve the performance of knowledge workers. For example, financial analysts, attorneys, business-school professors, market makers, portfolio managers, reporters, and many others would benefit from timely attention to certain business news stories. Bloomberg, Reuters, Bridge, and several other companies have profited greatly selling a variety of instant-access, business information services. In this project, we are investigating the use of inductive algorithms to build triage models. News-story importance can be assessed in many ways; here we concentrate on the use of prospective, objective criteria that can be assessed in many ways; here we concentrate on the prospective, objective criteria that can be used to build massive training sets automatically from historical data. Examples include the volume of follow-up stories and the stock-market reaction to a story. In both cases, for timely attention, it would be useful to be able to rank any given set of stories by expected importance. However, in order for the models to be used, it also is important to be able to understand what knowledge the models capture. Finally, the knowledge captured may be interesting in its own right; for example, business-school researchers are interested in firm-specific news that affects stock price. In this talk, I will discuss these problems and will present some preliminary results.

<u>Summary Notes on Workshop Presentation*:*</u>

An important problem is to monitor things, to perform triage and to identify patterns, so as to discover breaking news, to incorporate them in time, and to be able to recognize if they are important.

There are several people who need timely information, including professionals, business-school professors, attorneys, executives, reporters, librarians, financial and business people, etc.

A first question that can be raised is: what clues might exist for deciding a story's importance? There are some answers from the historical data, including: many similar stories follow immediately thereafter; similar articles have been important in the past; or, a combination of such information.

A general schema for the distribution of new stories to the users is: From various sources of information repository and information feeds, the information goes to the information monitoring and triage module, and from there, the selected information is distributed to the users.

As an example, consider the following question: Is a certain story important for the stock market? This is interpreted as "there is an increase of more than 1SD in the motion of the stock price, in the hour following the time the story appeared". Among all the distributed stories, only 9% meet this criterion.

Analyzing the ROC curves for Stock movement, for true positives versus false positives, created with random, TFIDF and Naïve Bayes classification, respectively, it was remarked that random classification was the worst, and TFIDF the best.

Regarding knowledge discovery in a model, there are specific questions, e.g., are we able to understand the knowledge in the model? How to understand knowledge in holistic models?

For example, consider just the high information gain words for the stock movement. This is not enough to understand the knowledge (looking at exemplar stories to identify artifacts is not a solid indicator; trying rule learning; in this respect, the holistic models perform better, but are still not enough).

In order to improve the methods for knowledge discovery, extract structured knowledge from holistic models, relabel data, and extract rules.

Yet this is not enough. A further improvement is based on intuition. For example, look at firm, specific news in history, and verify the semantics of the extracted rules. The results seem to be promising for detecting knowledge.

Further research directions include large-scale mining of multi-type data; concept drift; the possibility of this infrastructure to support other disciplines.

Integrating Background Information In Text Classification, Haym Hirsh, Rutgers University
(Joint work with S. Zelikovitz, Rutgers University)

Abstract

One popular class of supervised learning problems concerns the task of classifying items that are comprised primarily of text. Such problems can occur when assessing the relevance of a Web page to a given user, assigning topic labels to technical papers, and assessing the importance of a user's email. Most commonly this task is performed by extrapolating from a corpus of labeled textual training documents procedures for classifying further unlabeled documents. However, the widespread availability of enormous amounts of information in online form opens up this task to methods that can exploit additional sources information during classification. This talk will describe two techniques for doing so.

The first uses collections of background text into the vocabulary re-expression process performed by latent semantic indexing. We show that these two methods for integrating background information into text classification improves overall classification accuracy, particularly in cases where there is only modest amounts of data, as is the case when data must be obtained by hand-labeling.

Summary Notes on Workshop Presentation*:*

The text classifier problem is very important in supervised learning. There are several examples where text classification is performed: determining e-mail category, news story topic labeling, technical papers cataloging, web pages labeling, finding the industry segment corresponding to a certain corporate name, etc.

Among the methods used for solving the text classification problem there are:

- Traditional approach: human labeling. As background information it can be used on:
    - unlabeled data (learn on unlabeled data, then create classifiers etc.)
    - test data, and
    - other text artifacts in the same domain

- Using background text to assess document similarity

- Using background text with Latent Semantic Indexing
-
- Using numerical features in text classification

- Second order similarity methods
    - find background text items that is similar to the test example (using TFIDF cosine distance)
    - find background text items and for each item combine the scores using "noisy or" (this method performed efficiently using WHIRL)
- Strawman: Bootstrapping method; it consists in labeling each background item, adding it to the training data, then using the traditional validation technique

The Second order similarity method (WHIRL) and the Bootstrapping method were applied to find the industry segment corresponding to a corporate name on 4 datasets and applying 5 fold cross-validation. The second order similarity method performed the best with small datasets.

The Latent Semantic Indexing (LSI) for retrieval data is based on the following idea: create a matrix such that each word corresponds to a row and each column corresponds to a document. Then create data spaces throwing away spurious information, and assess similarity in the new space.

LSI using background text consists in the following: perform SVD on background text, then do LSI on background text, and use the background information to classify new items in a digital library.

In conclusion, the prevalence of online information should make us look for more information that can help with text classification. Approaches: Assign labels to the background text.

The punch line: Use the background text for retrieval. Let the sources of information drive the work.

PIRCS - An Effective Document Detection Tool, Kui Lam Kwok, Queens College

Abstract

PIRCS is an in-house developed IR engine and stands for Probabilistic Indexing and Retrieval - Components - System. It has been used in various forms to participate in all the past 9 TREC blind retrieval experiments with consistent and surprisingly good results. This talk will discuss the main ideas behind its approach, relationship with other models, a summary of its technology and samples of TREC experimental results for illustration.

Summary Notes on Workshop Presentation*:*

The general schema for extracting information from text is: from a large data set a smaller subset is determined by applying a filter. The resulting smaller data set is used for extracting and summarizing information, determining relationships, etc.

PIRCS (Probabilistic Indexing & Retrieval Components) has the role of a filter for the large data sets.

The probabilistic retrieval, introduced by Robertson And Sparck-Jones in 1976, solves the following problem: Given a query q, rank the documents optimally via the retrieval status value (RSV). In general, the following assumptions are made: there exist sample sets, and the documents are represented by binary vectors.

The Probabilistic Retrieval Model creates weights for a given query with irrelevant/relevant sets, but it does not provide irrelevant/relevant sets to start (bootstrap problem), nor does it handle nonbinary documents.

The PIRCS model assumptions are the following:

- explicit need statements (the queries must be of the same type as the documents)
- the queries and the documents are composed of independent conceptual components
- the queries and the documents are individually self-relevant.

This provides the basis for establishing relations among the components (for example words) which can then spread to other parts of the collection.

PIRCS provides the self-weighting of any document or any query using self-relevant sets, i.e., provides P(doc/query|relevant_doc/query) and P(doc/query|irrelevant_doc/query).

The advantages of PIRCS are the following: it can self-bootstrap, supports nonbinary documents, and allows symmetric RSV.

To become more stable, PIRCS uses a combination of RSV: RSV=αRSVq+(1-α)RSVd. In practice, α=0.5.

The Unigram Language Model is based on the following assumptions:

- the document/query are explicit and of the same type
- the documents are independent, nonbinary
- RSV is the probability of document generating a query (not vice-versa) and it is computed as RSV=log P(query(q1,…,qm)|doc).

We remark that this formula is similar to the formula for PIRCS's RSVd, except for the parameters. It may be concluded that PIRCS is in some sense a linear combination of probabilistic and linguistic approaches.

A new method of learning consists in learning from top-ranked pseudo-relevant documents in 2-stage retrieval, and it performs 50%-100% better than simple retrieval procedures.

A brief summary of the techniques in PIRCS includes: subdocument processing, term weighting, 2-word phrases, retrieval combination, pseudo-relevance feedback in 2-stage retrieval, collection enrichment, translation for English-Chinese CLIR.

Some directions for future work are: document retrieving, query-depending selective preprocessing, bigram term-dependence.

Searching The Web And Other Highly Diverse Document Collections: Some Practical And Theoretical Issues, Andras Kornai, Northern Light

Summary Notes on Workshop Presentation*:*

The plan of the talk is the following: first the focus will concentrate on the web and special collections. Then, on self-organization as unsupervised clustering. Finally, there will be presented the results of several experiments, and some open problems.

Northern Light (NL) deals with more than 350M pages (which represents less than 25% of Google's claimed activity, but rather more than that, of what is really on the Web). The special collections are represented by over 30M documents over 8K sources. NL performs subject classification, aka TDT, using NL's hierarchy of 16K subjects. NL does not provide deeper syntactic analysis, or translations into conceptual structures.

Self-organization refers to an algorithm that dynamically discovers the structure of the input, analyses the structure of the query, and adjusts the presentation of the output. In a "working" definition, it is an algorithm that discovers clusters.

Unsupervised clustering can be realized with or without non-random seeds, or with or without a preset number of clusters. It is known from our own experience that unsupervised clustering does not work.

The so-called dendogram trick presents a cluster tree that indicates the pedigree of sets; it is easy to compute, and gives a good graphical image. However, it is not practical.

Supervised clustering (e.g., k-mean clustering) is a great benefit for both speed and OCR (for example, refining acoustic models by bootstrapping). The method works as follows:

0. Obtain few examples where the truth is known
1. Build models on truthed examples.
2. Use models to obtain a larger/better set of truthed examples
3. Re-estimate models on larger/better set.
4. Repeat 2 and 3 until no improvement is seen.

There are two important remarks regarding supervised clustering: the seed data is not important, and the structure of the clusters is known in advance.

Language identification is one of the practical problems for which NL provides solutions.

In Information Retrieval (IR), the structure (identity) of clusters is not obvious, even if the truth is known. In this case, bootstrapping is useful.

For example, for the language interpretation, we know the sets of languages we are looking for, and we start from those with a bigger likelihood (the truth is the length of a page). We apply the supervised clustering method described above. We found that bootstrap helps, but expertise helps more (provides an improvement of 50-75%).

NL indexes the web along:

	range of an attribute	entropy
website, source	10^2 –10^8	16.25
primary subject	10^4	3.60
second subject	10^4	3.28
type	10^2	0.81
language	10^2	1.40

NL indexes over 373m documents. Skimming off the top, after removing 8m types, only 6.6% from the original (about 30m) documents are left. The website and the source, as well as the source and the language are almost "independent". The skimming procedure is not too harsh, as tested in practical experimentation.

Open problems and conclusions:

- Even the supervisible instances (length, type) are hard.
- The structure of the domain is ill understood.
- The amount of time we can spend analyzing a page is very limited.
- The automatic deduction of the structure is beyond the state of the art.

D. DEFINING A RESEARCH AGENDA

A panel discussion about future research needs led to a variety of recommendations.

The findings can be summarized in terms of (1) Research issues, (2) Technologies and (3) Development paths.

We found that the “bottom line” for **research** is: we must expand our use of additional data (queries, metadata, user data, ...) to develop automatic classification and analysis for digital libraries. No specific methods have priority in this, and many avenues must be explored in parallel.

The two key supporting **technologies** are (1) large sets of judged data that can be used for testing and (2) robust technologies to bring all sorts of data into digital and analyzable form.

A good **development path** is to embed these research and technology problems in a mission-relevant task such as (a) migration of technical reports and manuals or (b) integration of text and data libraries for health care support.

D1. An overall framework

To help define an agenda, the working group meeting included a one-hour session on “blue sky” efforts to define key research issues of importance to the nation and to the sponsoring organizations. Most of the ideas presented were given in their most general form and in what follows we refocus them to match an agency mission

All the workshop participants answered the questions presented below. The individual answers are summarized in Sections D2-D5 below. Note that the individual suggestions cannot be taken to represent a consensus, and the cost figures, in particular, may not reflect considered judgment. When a suggestion was made only in writing during the session, we do not know who provided it.

Charge to the workshop participants"

"Please fill in the slots P, M, D, Y in the following advice to the government of the United States

Dear Government,

In the area of Data Analysis and Digital Libraries, attention should focus on problem P, using method M, because expenditure of only D dollars should produce usable results in Y years."

In his talk, Paul Kantor (Rutgers) had described two gold standards against which algorithmic development must be measured. The first gold standard is the activity of a human librarian, who examines materials contained in a library and classifies them, thus resolving the collection into a set of overlapping sub-collections. The second gold standard applies to more narrowly focused sub-collections and is represented by the human analyst or scholar who is sent to that collection with the question "is there anything interesting here?". The words "anything interesting" can range from confirmation of existing theories to refutation of existing theories, the discovery of new topics, the formation of summaries of parts or the whole of the collection, and the discovery of contradictions, anomalies or inconsistencies. Self-structuration is the (hypothetical) process by which a digital library comes to understand its contents with reference to these two gold standards. It provides a context for other, more specific problems which can serve to focus mission oriented research over the next 3-5 years.

D2. Major Thrust: Comprehensive Medical Libraries

Elizabeth Liddy (Syracuse) spoke of the potentially enormous importance of combining full text library management with the fielded data contained in medical records. Presumably one of the functions of a health provider organization is to work from the set of symptoms into knowledge of the literature and to provide relevant documents ranging from highly specific treatment guidelines through less specific diagnostic flow charts, and finally to cutting edge research which may point the direction to economically or medically important new alternatives in diagnosis and treatment. This forms a very rich arena for studying linked collections of fielded data and digital libraries. All of the points made during the talk on the importance of dealing with multiple kinds of data in order to improve structuration and retrieval will thus apply. While Prof. Liddy presented the idea in the context of a national focus on health, it seems clearly to also have specific implications for all branches of the armed services, as they are responsible for maintaining the health of millions of people of all ages and exposed to a bewildering variety of unfamiliar or hostile environments.

D2.1. Data Mining of Medical Facts

Data mining of medical facts about a person from all of that person's medical records could formulate a query against the medical literature to inform people of health risks and treatment alternatives. The methods would include data mining, text mining and information retrieval and this would be a very large project scoped at perhaps 100 million dollars over 10 years.

D3. Major Thrust: Training Sets

Warren Greiff (MITRE) spoke of the importance of having large judged sets of data for any task for which we want to develop algorithmic methods. If we want to have algorithms that can find patterns in digital libraries such as internal substructure collection, hierarchical relationships among documents, contradictory relations and so forth, we need substantial non-trivial judged sets that can be used for the evaluation and development of algorithms.

D3.1. Metrics for Use in Offline Research

An important problem is determining metrics which can be used offline for inexpensive calibration and direction-setting during research projects. No cost or time frame was proposed for this activity.

D4. Major Thrust: Self-Structuration of Digital Libraries

D4.1. Unsupervised Learning and its Limits

Andras Kornai (Northern Light) reported that unstructured learning in general seems to be extremely difficult to do. Reporting on internal research at Northern Light, he suggested that it is hard to get to say 90% accuracy in the apparently trivial task of deciding the language of a web site. Thus he said that emphasis should be placed on developing good methods for supervised learning.

D4.2. Autoabstracting

(Anonymous suggestion). Automatic abstracting and summarization of text corpora, using statistical NLP and other processes as they work. Estimated budget: 10 billion dollars for visible results in two years.

D4.3. Customization of Results

(Anonymous suggestion). Customization of search or information retrieval to include information about a user's context broadly defined. Usable results should be available within one year.

D4.4. Semantics and Discourse for Web Documents

(Anonymous suggestion).Semantic and discourse analysis of text as applied particularly to web documents. The method to be used would be natural language processing. It's projected that some significant results could be obtained in four years for 20 million dollars.

D4.5. Role of Context in Information Use

(Anonymous suggestion). Better understanding of the role of context in information use. Methods formed by combining artificial intelligence in information retrieval. Results in about three years.

D4.6. Reduction of Biases by Analyzing Whole Web

(Anonymous suggestion). Applying information retrieval to both the public and the hidden web to reduce existing biases in what is presented. Methods are not yet specified but 3 million dollars in three years should produce results.

D4.7. Large Collections for Background Support

(Anonymous suggestion). Using existing large document collections to provide background or complementary knowledge to support data analysis. The methods will come from information retrieval, artificial intelligence and knowledge and data discovery. Projected to be able to produce results in five years at 100 million dollars.

D4.8. Automated Discovery of Topic areas

Paul Kantor (Rutgers) suggested automated structuration or mapping of collections into topic areas, based on using data from user judgments. Estimated to produce meaningful results in three to four years for 2 million dollars.

D4.9. Detecting Contradiction in Digital Libraries

(Anonymous suggestion). Detecting contradictions in digital libraries. The methods to be used involve machine learning on human judged training sets. It should be possible to make substantial progress in five years for one or two million dollars.

(Anonymous suggestion). Digitizing printed and written data. The method is to develop much better scanning infrastructure and some sort of distributed Napster-like storage and public domain OCR. Results should be achievable for 10 million dollars in three years.

D4.10. Text Retrieval and Categorization

(Anonymous suggestion). Text retrieval and categorization in filtering using improved feature selection with simple classifiers such as naïve Bayes. Improvements should be achievable in three years for 3 million dollars. This has been an important topic in the TREC setting, but is to some extent lost among the other tracks at TREC.

D4.11. Automatic Learning of Better Query Formation

(Anonymous suggestion). Developing techniques to learn how to create good queries for retrieval by learning from the structure and performance of prior queries. A variety of methods should be applicable and it should be possible to produce significant results for 5 million dollars in two years.

D5. Transformation across Media

Andras Kornai (Northern Light) discussed the importance of data entry for legacy collections. This can be approached by working for improvements in multi-faceted models for optical character recognition and automatic digitization of documents. While the commercial sector has largely moved the digitization process offshore, this has in effect dampened development of automated OCR techniques. It has been an economically feasible solution to accept the problems in existing OCR, and to hire inexpensive offshore labor to perform the final changes.

However, this is a narrow and self-limited view. First, it focuses attention on English language texts. For these texts it is relatively easy to find large populations of skilled workers who can process the texts, and produce results which can be screened for quality by on-shore workers. However, a large body of material in graphical form (maintenance diagrams, circuit diagrams, and so forth) is not a good candidate for this process. The skills base for manual editing of such materials is not widespread and not likely to become widespread.

In addition, from a mission perspective, large quantities of the legacy material of importance to the defense establishment represent sensitive material which cannot reasonably be sent out of the country for processing. Anyone who has worked with the technical reports, dating back to the 1950’s, which remain important in present-day work on missile defense systems, realizes the severity of this problem.

D5.1. Transcoding

(Anonymous suggestion). Transcoding to bring materials from one display medium to another. No methods or budget proposed but anticipated that results should be visible in about two years.

D5.2. Visually Impaired

(Anonymous suggestion). Tools for Search and Information Display for the Visually Impaired. It should be possible to produce audible results in two years.

D5.3. Filtering Multimedia

(Anonymous suggestion). Filtering of multi-media documents. The techniques to be used are seen as statistical. Five years and 100 million dollars.

D6. Summary of Discussion

We summarize the four broad themes that emerged as challenge areas for advancing the general area of data analysis for digital libraries.

D6.1. Aspects of Self-Structuration

If there was a consensus at the working group, it was that we need to find ways to make greater use of other kinds of data (besides the data contained in the digital documents themselves) in order to advance the structuration and self-analysis task for digital libraries. The data that might be involved were described in many ways. Mark Hansen of Bell Labs described it as outside information, Haym Hirsh of Rutgers described it as background information, and Paul Kantor of Rutgers has described it in other arenas as exogenous information. K.L. Kwok of Queens College gave a powerful specific example where additional data for the structuration of retrievable texts is built from a corpus of questions or problems for that text. This seems to point to an important direction for digital libraries, as the analysis of the content of digital libraries and of self-structuration will surely be helped by a better understanding of the set of problems which human scholars and analysts bring to the libraries. It appears that a broad range of techniques (statistical, pattern analysis, machine learning, etc.) all have something to contribute to this enterprise.

D6.2. Managing Data Across Media

Another important issue is the automatic migration of data from one format to another. This includes: print to electronic, and electronic to audible, as examples. One instance of this, the migration of old technical reports and maintenance manuals, is clearly of interest to mission agencies.

D6.3. Corpora of Training Data

For every type of task, bodies of tagged, judged materials are essential so that researchers can explore multiple alternative methods in a laboratory setting, with turnaround measured in hours or days, as opposed to the one year TREC cycle. An ideal set would consist of two or more parts, with the first part available for general experimentation, and the other parts available to provide an independent reality check on systems that appear to have achieved benchmark performance on the widely available part of the corpus.

D6.4. Comprehensive Model Problems

Finally, it appears desirable to search for mission-oriented comprehensive model problems, which will encompass all of the specific key research goals. Example model problems include the integration of health and medical records, and the automated management of libraries of manuals and technical reports now existing in paper formats only.