NIILM University

Information Retrival

Contents

Information Retrieval

Introduction

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Overview

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

History

“ But do you know that, although I have kept the diary [on a phonograph] for months past, it never once struck me how I was going to find any particular part of it in case I wanted to look it up? ”

—Dr Seward, Bram Stoker's Dracula, 1897

The idea of using computers to search for relevant pieces of information was popularized in the article As We May Think by Vannevar Bush in 1945. The first automated information retrieval systems were introduced in the 1950s and 1960s. By 1970 several different techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand

documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

Model types

For effectively retrieving relevant documents by IR strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporate a specific model for its document representation purposes. The picture on the right illustrates the relationship of some common models. In the picture, the models are categorized according to two dimensions: the mathematical basis and the properties of the model.

First dimension: mathematical basis

- Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:
- Standard Boolean model
- Extended Boolean model
- Fuzzy retrieval
- Algebraic models represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.
- Vector space model
- Generalized vector space model
- (Enhanced) Topic-based Vector Space Model
- Extended Boolean model
- Latent semantic indexing aka latent semantic analysis

- Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Bayes' theorem are often used in these models.
- Binary Independence Model
- Probabilistic relevance model on which is based the okapi (BM25) relevance function
- Uncertain inference
- Language models
- Divergence-from-randomness model
- Latent Dirichlet allocation
- Feature-based retrieval models view documents as vectors of values of feature functions (or just features) and seek the best way to combine these features into a single relevance score, typically by learning to rank methods. Feature functions are arbitrary functions of document and query, and as such can easily incorporate almost any other retrieval model as just a yet another feature.

Second dimension: properties of the model

- Models without term-interdependencies treat different terms/words as independent. This fact is usually represented in vector space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables.
- Models with immanent term interdependencies allow a representation of interdependencies between terms. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. by dimensional reduction) from the co-occurrence of those terms in the whole set of documents.
- Models with transcendent term interdependencies allow a representation of interdependencies between terms, but they do not allege how the interdependency between two terms is defined. They relay an external source for the degree of interdependency between two terms. (For example a human or sophisticated algorithms.)

Performance and correctness measures

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to

be either relevant or non-relevant to a particular query. In practice queries may be ill-posed and there may be different shades of relevancy.

Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

In binary classification, precision is analogous to positive predictive value. Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.

Note that the meaning and usage of "precision" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and statistics.

Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

In binary classification, recall is often called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

Fall-out

The proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

In binary classification, fall-out is closely related to specificity and is equal to. It can be looked at as the probability that a non-relevant document is retrieved by the query.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-measure

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

This is also known as the measure, because recall and precision are evenly weighted.

The general formula for non-negative real is:

Two other commonly used F measures are the measure, which weights recall twice as much as precision, and the measure, which weights precision twice as much as recall.

The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure . Their relationship is where .

Average precision

Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, one can plot a precision-recall curve, plotting precision as a function of recall . Average precision computes the average value of over the interval from to :

That is the area under the precision-recall curve. This integral is in practice replaced with a finite sum over every position in the ranked sequence of documents:

where is the rank in the sequence of retrieved documents, is the number of retrieved documents, is the precision at cut-off in the list, and is the change in recall from items to .

This finite sum is equivalent to:

where is an indicator function equaling 1 if the item at rank is a relevant document, zero otherwise. Note that the average is over all relevant documents and the relevant documents not retrieved get a precision score of zero.

Some authors choose to interpolate the function to reduce the impact of "wiggles" in the curve. For example, the PASCAL Visual Object Classes challenge (a benchmark for computer vision object detection) computes average precision by averaging the precision over a set of evenly spaced recall levels {0, 0.1, 0.2, ... 1.0}:

where is an interpolated precision that takes the maximum precision over all recalls greater than

An alternative is to derive an analytical function by assuming a particular parametric distribution for the underlying decision values. For example, a binormal precision-recall curve can be obtained by assuming decision values in both classes to follow a Gaussian distribution.

R-Precision

Precision at R-th position in the ranking of results for a query that has R relevant documents. This measure is highly correlated to Average Precision. Also, Precision is equal to Recall at the R-th position.

Mean average precision

Mean average precision for a set of queries is the mean of the average precision scores for each query.

where Q is the number of queries.

Discounted cumulative gain

Main article: Discounted cumulative gain

DCG uses a graded relevance scale of documents from the result set to evaluate the usefulness, or gain, of a document based on its position in the result list. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.

The DCG accumulated at a particular rank position is defined as:

Since result set may vary in size among different queries or systems, to compare performances the normalised version of DCG uses an ideal DCG. To this end, it sorts documents of a result list by relevance, producing an ideal DCG at position p ( ), which normalizes the score:

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. Note that in a perfect ranking algorithm, the will be the same as the producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable.

Other Measures

- Mean reciprocal rank
- Spearman's rank correlation coefficient

Collaborative information seeking

Collaborative information seeking (CIS) is a field of research that involves studying situations, motivations, and methods for people working in collaborative groups for information seeking projects, as well as building systems for supporting such activities. Such projects often involve information searching or information retrieval (IR), information gathering, and information sharing. Beyond that, CIS can extend to collaborative information synthesis and collaborative sense-making.

Background

Seeking for information is often considered a solo activity, but there are many situations that call for people working together for information seeking. Such situations are typically complex in nature, and involve working through several sessions exploring, evaluating, and gathering relevant information. Take for example, a couple going on a trip. They have the same goal, and in order to accomplish their goal, they need to seek out several kinds of information, including flights, hotels, and sightseeing. This may involve them working together over multiple sessions, exploring and collecting useful information, and collectively making decisions that help them move toward their common goal.

It is a common knowledge that collaboration is either necessary or highly desired in many activities that are complex or difficult to deal with for an individual. Despite its natural appeal and situational necessity, collaboration in information seeking is an understudied domain. The nature of the available information and its role in our lives have changed significantly, but the methods and tools that are used to access and share that information in collaboration have remained largely unaltered. People still use general-purpose systems such as email and IM for doing CIS projects, and there is a lack of specialized tools and techniques to support CIS explicitly.

There are also several models to explain information seeking and information behavior, but the areas of collaborative information seeking and collaborative information behavior remain understudied. A few specialized systems for supporting CIS have emerged in the recent past, but their usage and evaluations have underwhelmed. Despite such limitations, the field of CIS has been getting a lot of attention lately, and several promising theories and tools have come forth. A recent review of CIS related literature is written by Shah. His new book on this topic provides a comprehensive review of this field, including theories, models, systems, evaluation, and future research directions. Other notable books in this area include one by Morris and Teevan, as well as Foster's book on collaborative information behavior.

Theories

Depending upon what one includes or excludes while talking about CIS, we have many or hardly any theories. If we consider the past work on the groupware systems, many interesting insights can be obtained about people working on collaborative projects, the issues they face, and the guidelines for system designers. One of the notable works is by Grudin, who laid out eight design principles for developers of groupware systems.

The discussion below is primarily based on some of the recent works in the field of computer supported cooperative work CSCW, collaborative IR, and CIS.

Definitions and terminology

The literature is filled with works that use terms such as collaborative information retrieval, social searching, concurrent search, collaborative exploratory search, co-browsing, collaborative information behavior, collaborative information synthesis, and collaborative information seeking, which are often used interchangeably.

There are several definitions of such related or similar terms in the literature. For instance, Foster defined collaborative IR as "the study of the systems and practices that enable individuals to collaborate during the seeking, searching, and retrieval of information." Shah defined CIS as a process of collaboratively seeking information that is "defined explicitly among the participants, interactive, and mutually beneficial." While there is still a lack of a definition or a terminology that is universally accepted, but most agree that CIS is an active process, as opposed to collaborative filtering, where a system connects the users based on their passive involvement (e.g., buying similar products on Amazon).

Models of collaboration

Foley and Smeaton defined two key aspects of collaborative information seeking as division of labor and the sharing of knowledge. Division of labor allows collaborating searchers to tackle larger problems by reducing the duplication of effort (e.g., finding documents that one's collaborator has already discovered). The sharing of knowledge allows searchers to influence each other's activities as they interact with the retrieval system in pursuit of their (often evolving) information need. This influence can occur in real time if the collaborative search system supports it, or it can occur in a turn-taking, asynchronous manner if that is how interaction is structured.

Teevan et al. characterized two classes of collaboration, task-based vs. trait-based. Task-based collaboration corresponds to intentional collaboration; trait-based collaboration facilitates the sharing of knowledge through inferred similarity of information need.

Situations, motivations, and methods for CIS

One of the important issues to study in CIS is the instance, reason, and the methods behind a collaboration. For instance, Morris, using a survey with 204 knowledge workers at a large technology company found that people often like and want to collaborate, but they do not find specialized tools to help them in such endeavors. Some of the situations for doing collaborative information seeking in this survey were travel planning, shopping, and literature search. Shah,[22] similarly, using personal interviews, identified three main reasons why people collaborate.

1. Requirement/setup. Sometimes a group of people are "forced" to collaborate. Example includes a merger between two companies.

2. Division of labor. Working together may help the participants to distribute the workload. Example includes a group of students working on a class project.

3. Diversity of skills. Often people get together because they could not individually possess the required set of skills. Example includes co-authorship, where different authors bring different set of skills to the table.

As far as the tools and/or methods for CIS are concerned, both Morris and Shah found that email is still the most used tool. Other popular methods are face-to-face meetings, IM, and phone or conference calls. In general, the choice of the method or tool for our respondents depended on their situation (co-located or remote), and objective (brainstorming or working on independent parts).

Space-time organization of CIS systems and methods

The classical way of organizing collaborative activities is based on two factors: location and time. Recently Hansen & Jarvelin and Golovchinsky, Pickens, & Back also classified approaches to collaborative IR using these two dimensions of space and time. See "Browsing is a Collaborative Process", where the authors depict various library activities on these two dimensions.

As we can see from this figure, the majority of collaborative activities in conventional libraries are co-located and synchronous, whereas collaborative activities relating to digital libraries are more remote and synchronous. Social information filtering, or collaborative filtering, as we saw earlier, is a process benefitting from other users' actions in the past; thus, it falls under asynchronous and mostly remote domain. These days email also serves as a tool for doing asynchronous collaboration among users who are not co-located. Chat or IM (represented as `internet' in the figure) helps to carry out synchronous and remote collaboration.

Rodden, similarly, presented a classification of CSCW systems using the form of interaction and the geographical nature of cooperative systems. Further, Rodden & Blair presented an important characteristic to all CSCW systems - control. According to the authors, two predominant control mechanisms have emerged within CSCW systems: speech act theory systems, and procedure based systems. These mechanisms are tightly coupled with the kind of control the system can support in a collaborative environment (discussed later).

Often researchers also talk about other dimensions, such as intentionality and depth of mediation (system mediated or user mediated), while classifying various CIS systems.

Control, communication, and awareness

Three components specific to group-work or collaboration that are highly predominant in the CIS or CSCW literature are control, communication, and awareness. In this section key definitions and related works for these components will be highlighted. Understanding their roles can also help us address various design issues with CIS systems.

Control

Rodden identified the value of control in CSCW systems and listed a number of projects with their corresponding schemes for implementing for control. For instance, the COSMOS project had a formal structure to represent control in the system. They used roles to represent people or automatons, and rules to represent the flow and processes. Roles of the people could be supervisor, processor, or analyst. Rules could be a condition that a process needs to satisfy in order to start or finish. Due to such a structure seen in projects like COSMOS, Rodden classified these control systems as procedural based systems.

Communication

This is one of the most critical components of any collaboration. In fact, Rodden (1991) identified message or communication systems as the class of systems in CSCW that is most mature and most widely used.

Since the focus here is on CIS systems that allow its participants to engage in an intentional and interactive collaboration, there must be a way for the participants to communicate with each other. What is interesting to note is that often, collaboration could begin by letting a group of users communicate with each other. For instance, Donath & Robertson presented a system that allows a user to know that others were currently viewing the same webpage and communicate with those people to initiate a possible collaboration or at least a co-browsing experience. Providing communication capabilities even in an environment that was not originally designed for carrying out collaboration is an interesting way of encouraging collaboration.

Awareness

Awareness, in the context of CSCW, has been defined as "an understanding of the activities of others, which provides a context for your own activity". The following four kinds of awareness are often discussed and addressed in the CSCW literature:

1. Group awareness. This kind of awareness includes providing information to each group member about the status and activities of the other collaborators at a given time.

2. Workspace awareness. This refers to a common workspace that the group has where they can bring and discuss their findings, and create a common product.

3. Contextual awareness. This type of awareness relates to the application domain, rather than the users. Here, we want to identify what content is useful for the group, and what the goals are for the current project.

4. Peripheral awareness. This relates to the kind of information that has resulted from personal and the group's collective history, and should be kept separate from what a participant is currently viewing or doing.

Shah and Marchionini studied awareness as provided by interface in collaborative information seeking. They found that one needs to provide "right" (not too little, not too much, and appropriate for the task at hand) kind of awareness to reduce the cost of coordination and maximize the benefits of collaboration.

Systems

A number of specialized systems have been developed back from the days of the groupware systems to today's Web 2.0 interfaces. A few such examples, in chronological order, are given below.

Ariadne

Twidale et al. developed Ariadne to support the collaborative learning of database browsing skills. In addition to enhancing the opportunities and effectiveness of the collaborative learning that already occurred, Ariadne was designed to provide the facilities that would allow collaborations to persist as people increasingly searched information remotely and had less opportunity for spontaneous face-to-face collaboration.

Ariadne was developed in the days when Telnet-based access to library catalogs was a common practice. Building on top of this command-line interface, Ariadne could capture the users' input and the database's output, and form them into a search history that consisted of a series of command-output pairs. Such a separation of capture and display allowed Ariadne to work with various forms of data capture methods.

To support complex browsing processes in collaboration, Ariadne presented a visualization of the search process. This visualization consisted of thumbnails of screens, looking like playing cards, which represented command-output pairs. Any such card can be expanded to reveal its details. The horizontal axis on Ariadne's display represented time, and the vertical axis showed information on the semantics of the action it represented: the top row for the top level menus, the middle row for specifying a search, and the bottom row for looking at particular book details.

This visualization of the search process in Ariadne makes it possible to annotate, discuss with colleagues around the screen, and distribute to remote collaborators for asynchronous commenting easily and effectively. As we saw in the previous section, having access to one's history as well as the history of one's collaborators are very crucial to effective collaboration.

Ariadne implements these requirements with the features that let one visualize, save, and share a search process. In fact, the authors found one of the advantages of search visualization was the ability to recap previous searching sessions easily in a multi-session exploratory searching.

Search Together

More recently, one of the collaborative information seeking tools that have caught a lot of attention is Search Together, developed by Morris and Horvitz. The design of this tool was motivated by a survey that the researchers did with 204 knowledge workers, in which they discovered the following.

- A majority of respondents wanted to collaborate while searching on the Web.

- The most common ways of collaborating in information seeking tasks are sending emails back and forth, using IM to exchange links and query terms, and using phone calls while looking at a Web browser.

- Some of the most popular Web searching tasks on which people like to collaborate are planning travels or social events, making expensive purchases, researching medical conditions, and looking for information related to a common project.

Based on the survey responses, and the current and desired practices for collaborative search, the authors of Search Together identified three key features for supporting people's collaborative information behavior while searching on the Web: awareness, division of labor, and persistence. Let us look at how these three features are implemented.

Search Together instantiates awareness in several ways, one of which is per-user query histories. This is done by showing each group member's screen name, his/her photo and queries in the "Query Awareness" region. The access to the query histories is immediate and interactive, as clicking on a query brings back the results of that query from when it was executed. The authors identified query awareness as a very important feature in collaborative searching, which allows group members to not only share their query terms, but also learn better query formulation techniques from one another.

Another component of SearchTogether that facilitates awareness is the display of page-specific metadata. This region includes several pieces of information about the displayed page, including group members who viewed the given page, and their comments and ratings. The authors claim that such visitation information can help one either choose to avoid a page already visited by someone in the group to reduce the duplication of efforts, or perhaps choose to visit such pages, as they provide a sign of promising leads as indicated by the presence of comments and/or ratings.

Division of labor in SearchTogether is implemented in three ways: "Split Search" allows one to split the search results among all online group members in a round-robin fashion,"Multi-Engine

Search" takes a query and runs it on n different search engines, where n is the number of online group members, manual division of labor can be facilitated using integrated IM.

Finally, the persistence feature in SearchTogether is instantiated by storing all the objects and actions, including IM conversations, query histories, recommendation queues, and page-specific metadata. Such data about all the group members are available to each member when he/she logs in. This allows one to easily carry a multi-session collaborative project.

Cerchiamo

Cerchiamo is a collaborative information seeking tool that explores issues related to algorithmic mediation of information seeking activities and how collaborators' roles can be used to structure the user interface. Cerchiamo introduced the notion of algorithmic mediation, that is, the ability of the system to collect input asynchronously from multiple collaborating searchers, and to use these multiple streams of input to affect the information that is being retrieved and displayed to the searchers.

Cerchiamo collected judgments of relevance from multiple collaborating searchers and used those judgments to create a ranked list of items that were potentially relevant to the information need. This algorithm prioritized items that were retrieved by multiple queries and that were retrieved by queries that also retrieved many other relevant documents. This rank fusion is just one way in which a search system that manages activities of multiple collaborating searchers can combine their inputs to generate results that are better than those produced by individuals working independently.

Cerchiamo implemented two roles—Prospector and Miner—that searchers could assume. Each role had an associated interface. The Prospector role/interface focused on running many queries and making a few judgments of relevance for each query to explore the information space. The Miner role/interface focused on making relevance judgments on a ranked list of items selected from items retrieved by all queries in the current session. This combination of roles allowed searchers to explore and exploit the information space, and led teams to discover more unique relevant documents than pairs of individuals working separately.

Coagmento

Coagmento (Latin for "working together") is a new and unique system that allows a group of people work together for their information seeking tasks without leaving their browsers. Coagmento has been developed with a client-server architecture, where the client is implemented as a Firefox plug-in that helps multiple people working in collaboration to communicate, and search, share and organize information. The server component stores and provides all the objects and actions collected from the client. Due to this decoupling, Coagmento provides a flexible architecture that allows its users to be co-located or remote, working synchronously or asynchronously, and use different platforms.

Coagmento includes a toolbar and a sidebar. The toolbar has several buttons that helps one collect information and be aware of the progress in a given collaboration. The toolbar has three major parts:

- Buttons for collecting information and making annotations. These buttons help one save or remove a webpage, make annotations on a webpage, and highlight and collect text snippets.

- Page-specific statistics. The middle portion of the toolbar shows various statistics, such as the number of views, annotations, and snippets, for the displayed page. A user can click on a given statistic and obtain more information. For instance, clicking on the number of snippets will bring up a window that shows all the snippets collected by the collaborators from the displayed page.

- Project-specific statistics. The last portion of the toolbar displays task/project name and various statistics, including number of pages visited and saved, about the current project. Clicking on that portion brings up the workspace where one can view all the collected objects (pages and snippets) brought in by the collaborators for that project.

The sidebar features a chat window, under which there are three tabs with the history of search engine queries, saved pages and snippets. With each of these objects, the user who created or collected that object is shown. Anyone in the group can access an object by clicking on it. For instance, one can click on a query issued by anyone in the group to re-run that query and bring up the results in the main browser window.

An Android (operating system) app for Coagmento can be found in the Android Market.

Cosme

Fernandez-Luna et al. introduce Cosme (COde Search MEeting) as a NetBeans IDE plug-in that enables remote team of software developers to collaborate in real time during source-code search sessions. The COSME design was motivated by early stadies of C. Foley, M. R. Morris, C. Shah, among others researchers, and by habits of software developers identified in a survey of 117 universities students and professors related with projects of software development, as well as to computer programmers of some companies. The five more commons collaborative search habits (or related to it) of the interviewees was:

- Revision of problems by the team in the workstation of one of them.

- Suggest addresses of Web pages that they have already visited previously, digital books stored in some FTP, or source files of a version control system.

- Send emails with algorithms or explanatory text.

- Division of search tasks among each member of the team for sharing the final result.

- Store relevant information in individual workstation.

COSME is designed to enable either synchronous or asynchronous, but explicit remote collaboration among team developers with shared technical information needs. Its client user interface include a search panel that lets developers to specify queries, division of labor principle (possible combination include the use of different search engines, ranking fusion, and split algorithms), searching field (comments, source-code, class or methods declaration), and the collection type (source-code files or digital documentation). The sessions panel wraps the principal options to management the collaborative search sessions, which consists in a team of developers working together to satisfy their shared technical information needs. For example, a developer can use the embedded chat room to negotiate the creation of a collaborative search session, and show comments of the current and historical search results. The implementation of Cosme was based on CIRLab (Collaborative Information Retrieval Laboratory) instantiation, a groupware framework for CIS research and experimentation, Java as programming language, NetBeans IDE Platform as plug-in base, and Amenities (A MEthodology for aNalysis and desIgn of cooperaTIve systEmS) as software engineering methodology.

Open-source application frameworks and toolkits

CIS systems development is a complex task, which involves software technologies and Know-how in different areas such as distributed programming, information search and retrieval, collaboration among people, task coordination and many others according to the context. This situation is not ideal because it requires great programming efforts. Fortunately, some CIS application frameworks and toolkits are increasing their popularity since they have a high reusability impact for both developers and researchers, like Coagmento Collaboratory and Drakkar Keel.

Human–computer information retrieval

his term human–computer information retrieval was coined by Gary Marchionini in a series of lectures delivered between 2004 and 2006.[4] Marchionini's main thesis is that "HCIR aims to empower people to explore large-scale information bases but demands that people also take responsibility for this control by expending cognitive and physical energy."

In 1996 and 1998, a pair of workshops at the University of Glasgow on information retrieval and human–computer interaction sought to address the overlap between these two fields. Marchionini notes the impact of the World Wide Web and the sudden increase in information literacy – changes that were only embryonic in the late 1990s.

A few workshops have focused on the intersection of IR and HCI. The Workshop on Exploratory Search, initiated by the University of Maryland Human-Computer Interaction Lab in 2005,

alternates between the Association for Computing Machinery Special Interest Group on Information Retrieval (SIGIR) and Special Interest Group on Computer-Human Interaction (CHI) conferences. Also in 2005, the European Science Foundation held an Exploratory Workshop on Information Retrieval in Context. Then, the first Workshop on Human Computer Information Retrieval was held in 2007 at the Massachusetts Institute of Technology.

What is HCIR?

HCIR includes various aspects of IR and HCI. These include exploratory search, in which users generally combine querying and browsing strategies to foster learning and investigation; information retrieval in context (i.e., taking into account aspects of the user or environment that are typically not reflected in a query); and interactive information retrieval, which Peter Ingwersen defines as "the interactive communication processes that occur during the retrieval of information by involving all the major participants in information retrieval (IR), i.e. the user, the intermediary, and the IR system."

A key concern of HCIR is that IR systems intended for human users be implemented and evaluated in a way that reflects the needs of those users.

Most modern IR systems employ a ranked retrieval model, in which the documents are scored based on the probability of the document's relevance to the query. In this model, the system only presents the top-ranked documents to the user. This systems are typically evaluated based on their mean average precision over a set of benchmark queries from organizations like the Text Retrieval Conference (TREC).

Because of its emphasis in using human intelligence in the information retrieval process, HCIR requires different evaluation models – one that combines evaluation of the IR and HCI components of the system. A key area of research in HCIR involves evaluation of these systems. Early work on interactive information retrieval, such as Juergen Koenemann and Nicholas J. Belkin's 1996 study of different levels of interaction for automatic query reformulation, leverage the standard IR measures of precision and recall but apply them to the results of multiple iterations of user interaction, rather than to a single query response. Other HCIR research, such as Pia Borlund's IIR evaluation model, applies a methodology more reminiscent of HCI, focusing on the characteristics of users, the details of experimental design, etc.

Goals

Marchionini put forth the following goals towards a system where the user has more control in determining relevant results:

- Systems should aim to get people closer to the information they need, especially to the meaning; that is, systems can no longer only deliver the relevant documents, but must also provide facilities for making meaning with those documents.

- Systems should increase user responsibility as well as control; that is, information systems require human intellectual effort, and good effort is rewarded.

- Systems should have flexible architectures so they may evolve and adapt to increasingly more demanding and knowledgeable installed bases of users over time.

- Systems should aim to be part of information ecology of personal and shared memories and tools rather than discrete standalone services.

- Systems should support the entire information life cycle (from creation to preservation) rather than only the dissemination or use phase.

- Systems should support tuning by end users and especially by information professionals who add value to information resources.

- Systems should be engaging and fun to use.

In short, Marchionini seems to expect information retrieval systems to operate in the way that good libraries do. Systems should help users to bridge the gap between data or information (in the very narrow, granular sense of these terms) and knowledge (processed data or information that provides the context necessary to inform the next iteration of an information seeking process). That is, good libraries provide both the information a patron needs as well as a partner in the learning process-—the information professional—-to navigate that information, make sense of it, preserve it, and turn it into knowledge (which in turn creates new, more informed information needs .

Techniques

The techniques associated with HCIR emphasize representations of information that use human intelligence to lead the user to relevant results. These techniques also strive to allow users to explore and digest the dataset without penalty, i.e., without expending unnecessary costs of time, mouse clicks, or context shift.

Many search engines have features that incorporate HCIR techniques. Spelling suggestions and automatic query reformulation provide mechanisms for suggesting potential search paths that can lead the user to relevant results. These suggestions are presented to the user, putting control of selection and interpretation in the user's hands.

Faceted search enables users to navigate information hierarchically, going from a category to its sub-categories, but choosing the order in which the categories are presented. This contrasts with traditional taxonomies in which the hierarchy of categories is fixed and unchanging. Faceted

navigation, like taxonomic navigation, guides users by showing them available categories (or facets), but does not require them to browse through a hierarchy that may not precisely suit their needs or way of thinking.

Look ahead provides a general approach to penalty-free exploration. For example, various web applications employ AJAX to automatically complete query terms and suggest popular searches. Another common example of look ahead is the way in which search engines annotate results with summary information about those results, including both static information (e.g., metadata about the objects) and "snippets" of document text that are most pertinent to the words in the search query.

Relevance feedback allows users to guide an IR system by indicating whether particular results are more or less relevant.

Summarization and analytics help users digest the results that come back from the query. Summarization here is intended to encompass any means of aggregating or compressing the query results into a more human-consumable form. Faceted search, described above, is one such form of summarization. Another is clustering, which analyzes a set of documents by grouping similar or co-occurring documents or terms. Clustering allows the results to be partitioned into groups of related documents. For example, a search for "java" might return clusters for Java (programming language), Java (island), or Java (coffee).

Visual representation of data is also considered a key aspect of HCIR. The representation of summarization or analytics may be displayed as tables, charts, or summaries of aggregated data. Other kinds of information visualization that allow users access to summary views of search results include tag clouds and tree mapping.

Information extraction

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction.

Due to the difficulty of the problem, current approaches to IE focus on narrowly restricted domains. An example is the extraction from news wire reports of corporate mergers, such as denoted by the formal relation:from an online news sentence such as:

"Yesterday, New York based Foo Inc. announced their acquisition of Bar Corp."

A broad goal of IE is to allow computation to be done on the previously unstructured data. A more specific goal is to allow logical reasoning to draw inferences based on the logical content

of the input data. Structured data is semantically well-defined data from a chosen target domain, interpreted with respect to category and context.

History

Information extraction dates back to the late 1970s in the early days of NLP. An early commercial system from the mid-1980s was JASPER built for Reuters by the Carnegie Group with the aim of providing real-time financial news to financial traders.

Beginning in 1987, IE was spurred by a series of Message Understanding Conferences. MUC is a competition-based conference that focused on the following domains:

- MUC-1 (1987), MUC-2 (1989): Naval operations messages.
- MUC-3 (1991), MUC-4 (1992): Terrorism in Latin American countries.
- MUC-5 (1993): Joint ventures and microelectronics domain.
- MUC-6 (1995): News articles on management changes.
- MUC-7 (1998): Satellite launch reports.

Considerable support came from the U.S. Defense Advanced Research Projects Agency (DARPA), who wished to automate mundane tasks performed by government analysts, such as scanning newspapers for possible links to terrorism.

Present significance

The present significance of IE pertains to the growing amount of information available in unstructured form. Tim Berners-Lee, inventor of the world wide web, refers to the existing Internet as the web of documents [and advocates that more of the content be made available as a web of data. Until this transpires, the web largely consists of unstructured documents lacking semantic metadata. Knowledge contained within these documents can be made more accessible for machine processing by means of transformation into relational form, or by marking-up with XML tags. An intelligent agent monitoring a news data feed requires IE to transform unstructured data into something that can be reasoned with. A typical application of IE is to scan a set of documents written in a natural language and populate a database with the information extracted.

Tasks and subtasks

Applying information extraction on text, is linked to the problem of text simplification in order to create a structured view of the information present in free text. The overall goal being to create a more easily machine-readable text to process the sentences. Typical subtasks of IE include:

- Named entity extraction which could include:

o Named entity recognition: recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions, employing existing knowledge of the domain or information extracted from other sentences. Typically the recognition task involves assigning a unique identifier to the extracted entity. A simpler task is named entity detection, which aims to detect entities without having any existing knowledge about the entity instances. For example, in processing the sentence "M. Smith likes fishing", named entity detection would denote detecting that the phrase "M. Smith" does refer to a person, but without necessarily having (or using) any knowledge about a certain M. Smith who is (or, "might be") the specific person whom that sentence is talking about.

o Coreference resolution: detection of coreference and anaphoric links between text entities. In IE tasks, this is typically restricted to finding links between previously-extracted named entities. For example, "International Business Machines" and "IBM" refer to the same real-world entity. If we take the two sentences "M. Smith likes fishing. But he doesn't like biking", it would be beneficial to detect that "he" is referring to the previously detected person "M. Smith".

- Relationship extraction: identification of relations between entities, such as:
  PERSON works for ORGANIZATION (extracted from the sentence "Bill works for IBM.")
  PERSON located in LOCATION (extracted from the sentence "Bill is in France.")

  Semi-structured information extraction which may refer to any IE that tries to restore some kind information structure that has been lost through publication such as:
- Table extraction: finding and extracting tables from documents.
- Comments extraction : extracting comments from actual content of article in order to restore the link between author of each sentence
  Language and vocabulary analysis
- Terminology extraction: finding the relevant terms for a given corpus
  Audio extraction
- Template-based music extraction: finding relevant characteristic in an audio signal taken from a given repertoire; for instance time indexes of occurrences of percussive sounds can be extracted in order to represent the essential rhythmic component of a music piece.

Note this list is not exhaustive and that the exact meaning of IE activities is not commonly accepted and that many approaches combine multiple sub-tasks of IE in order to achieve a wider goal. Machine learning, statistical analysis and/or natural language processing are often used in IE.

IE on non-text documents is becoming an increasing topic in research and information extracted from multimedia documents can now be expressed in a high level structure as it is done on text. This naturally lead to the fusion of extracted information from multiple kind of documents and sources.

World Wide Web applications

IE has been the focus of the MUC conferences. The proliferation of the Web, however, intensified the need for developing IE systems that help people to cope with the enormous amount of data that is available online. Systems that perform IE from online text should meet the requirements of low cost, flexibility in development and easy adaptation to new domains. MUC systems fail to meet those criteria. Moreover, linguistic analysis performed for unstructured text does not exploit the HTML/XML tags and layout format that are available in online text. As a result, less linguistically intensive approaches have been developed for IE on the Web using wrappers, which are sets of highly accurate rules that extract a particular page's content. Manually developing wrappers has proved to be a time-consuming task, requiring a high level of expertise. Machine learning techniques, either supervised or unsupervised, have been used to induce such rules automatically.

Wrappers typically handle highly structured collections of web pages, such as product catalogues and telephone directories. They fail, however, when the text type is less structured, which is also common on the Web. Recent effort on adaptive information extraction motivates the development of IE systems that can handle different types of text, from well-structured to almost free text -where common wrappers fail- including mixed types. Such systems can exploit shallow natural language knowledge and thus can be also applied to less structured text.

Approaches

Three standard approaches are now widely accepted

- Hand-written regular expressions (perhaps stacked)
- Using classifiers
  - Generative: naïve Bayes classifier
  - Discriminative: maximum entropy models
- Sequence models

- Hidden Markov model
- CMMs/MEMMs
- Conditional random fields (CRF) are commonly used in conjunction with IE for tasks as varied as extracting information from research papers to extracting navigation instructions.

Numerous other approaches exist for IE including hybrid approaches that combine some of the standard approaches previously listed.

Data mining

Data mining (the analysis step of the "Knowledge Discovery and Data Mining" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Etymology

n the 1960s, statisticians used terms like "Data Fishing" or "Data Dredging" to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term "Data Mining" appeared around 1990 in the database community. At the beginning of the century, there was a phrase "database mining"™, trademarked by HNC, a San Diego-based company (now merged into FICO), to pitch their Data Mining Workstation; researchers consequently turned to "data mining". Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Piatetsky-Shapiro coined the term "Knowledge Discovery in Databases" for the first workshop on the same topic (1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

Research and evolution

The premier professional body in the field is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD). Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings, and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".

Computer science conferences on data mining include:

- CIKM Conference – ACM Conference on Information and Knowledge Management
- DMIN Conference – International Conference on Data Mining
- DMKD Conference – Research Issues on Data Mining and Knowledge Discovery
- ECDM Conference – European Conference on Data Mining
- ECML-PKDD Conference – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- EDM Conference – International Conference on Educational Data Mining
- ICDM Conference – IEEE International Conference on Data Mining
- KDD Conference – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- MLDM Conference – Machine Learning and Data Mining in Pattern Recognition
- PAKDD Conference – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining
- PAW Conference – Predictive Analytics World
- SDM Conference – SIAM International Conference on Data Mining (SIAM)
- SSTD Symposium – Symposium on Spatial and Temporal Databases
- WSDM Conference – ACM Conference on Web Search and Data Mining

Data mining topics are also present on many data management/database conferences such as the ICDE Conference, SIGMOD Conference and International Conference on Very Large Data Bases

Process

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection

(2) Pre-processing

(3) Transformation

(4) Data Mining

(5) Interpretation/Evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding

(2) Data Understanding

(3) Data Preparation

(4) Modeling

(5) Evaluation

(6) Deployment

or a simplified process such as

(1) pre-processing,

(2) data mining, and

(3) results validation.

Polls conducted in 2002, 2004, and 2007 show that the CRISP-DM methodology is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Data mining

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- Regression – attempts to find a function which models the data with the least error.

- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Results validation

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as overfitting, but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening.

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called over fitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves.

If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

Standards

There have been some efforts to define standards for the data mining process, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0). Development on successors to these processes (CRISP-DM 2.0 and JDM 2.0) was active in 2006, but has stalled since. JDM 2.0 was withdrawn without reaching a final draft.

For exchanging the extracted models – in particular for use in predictive analytics – the key standard is the Predictive Model Markup Language (PMML), which is an XML-based language developed by the Data Mining Group (DMG) and supported as exchange format by many data mining applications. As the name suggests, it only covers prediction models, a particular data mining task of high importance to business applications. However, extensions to cover (for example) subspace clustering have been proposed independently of the DMG.

Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called tablebases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully acquire the high level of abstraction required to be applied successfully. Instead, extensive experimentation with the tablebases – combined with an intensive study of tablebase-answers to well designed problems, and with knowledge of prior art (i.e., pre-tablebase knowledge) – is used to yield insightful patterns. Berlekamp (in dots-and-boxes, etc.) and John Nunn (in chess endgames) are notable examples of researchers doing this work, though they were not – and are not – involved in tablebase generation.

Business

Data mining is the analysis of historical business activities, stored as static data in data warehouse databases, to reveal hidden patterns and trends. Data mining software uses advanced pattern recognition algorithms to sift through large amounts of data to assist in discovering previously unknown strategic business information. Examples of what businesses use data mining for include performing market analysis to identify new product bundles, finding the root cause of manufacturing problems, to prevent customer attrition and acquire new customers, cross-sell to existing customers, and profile customers with more accuracy.

- In today's world raw data is being collected by companies at an exploding rate. For example, Walmart processes over 20 million point-of-sale transactions every day. This information is stored in a centralized database, but would be useless without some type of data mining software to analyse it. If Walmart analyzed their point-of-sale data with data mining techniques they would be able to determine sales trends, develop marketing campaigns, and more accurately predict customer loyalty.

- Every time a credit card or a store loyalty card is being used, or a warranty card is being filled, data is being collected about the users behavior. Many people find the amount of information stored about us from companies, such as Google, Facebook, and Amazon, disturbing and are concerned about privacy. Although there is the potential for our personal data to be used in harmful, or unwanted, ways it is also being used to make our lives better. For example, Ford

and Audi hope to one day collect information about customer driving patterns so they can recommend safer routes and warn drivers about dangerous road conditions.

- Data mining in customer relationship management applications can contribute significantly to the bottom line.[citation needed] Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which channel and to which offer an individual is most likely to respond (across all potential offers). Additionally, sophisticated applications could be used to automate mailing. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, "uplift modeling" can be used to determine which people have the greatest increase in response if given an offer. Uplift modeling thereby enables marketers to focus mailings and offers on persuadable people, and not to send offers to people who will buy the product without an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

- Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. Rather than using one model to predict how many customers will churn, a business could build a separate model for each region and customer type. Then, instead of sending an offer to all people that are likely to churn, it may only want to send offers to loyal customers. Finally, the business may want to determine which customers are going to be profitable over a certain window in time, and only send the offers to those that are likely to be profitable. In order to maintain this quantity of models, they need to manage model versions and move on to automated data mining.

- Data mining can be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Information obtained – such as universities attended by highly successful employees – can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

- Market basket analysis, relates to data-mining use in retail sales. If a clothing store records the purchases of customers, a data mining system could identify those customers who favor silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical, or inexact rules may also be present within a database.

- Market basket analysis has been used to identify the purchase patterns of the Alpha Consumer. Analyzing the data collected on this type of user has allowed companies to predict future buying trends and forecast supply demands.

- Data mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich database of history of their customer transactions for millions of customers dating back a number of years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

- Data mining for business applications can be integrated into a complex modeling and decision making process. Reactive business intelligence (RBI) advocates a "holistic" approach that integrates data mining, modeling, and interactive visualization into an end-to-end discovery and continuous innovation process powered by human and automated learning.

- In the area of decision making, the RBI approach has been used to mine knowledge that is progressively acquired from the decision maker, and then self-tune the decision method accordingly. The relation between the quality of a data mining system and the amount of investment that the decision maker is willing to make was formalized by providing an economic perspective on the value of “extracted knowledge” in terms of its payoff to the organization. This decision-theoretic classification framework was applied to a real-world semiconductor wafer manufacturing line, where decision rules for effectively monitoring and controlling the semiconductor wafer fabrication line were developed.

- An example of data mining related to an integrated-circuit (IC) production line is described in the paper "Mining IC Test Data to Optimize VLSI Testing." In this paper, the application of data mining and decision analysis to the problem of die-level functional testing is described. Experiments mentioned demonstrate the ability to apply a system of mining historical die-test data to create a probabilistic model of patterns of die failure. These patterns are then utilized to decide, in real time, which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products. Other examples of the application of data mining methodologies in semiconductor manufacturing environments suggest that data mining methodologies may be particularly useful when data is scarce, and the various physical and chemical parameters that affect the process exhibit highly complex interactions. Another implication is that on-line monitoring of the semiconductor manufacturing process using data mining may be highly effective.

- Ford and Audi hope to one day collect information about customer driving patterns so they can recommend safer routes and warn drivers about dangerous road conditions.

Science and engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

- In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases. One data mining method that is used to perform this task is known as multifactor dimensionality reduction.

- In the area of electrical power engineering, data mining methods have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on, for example, the status of the insulation (or other important safety-related parameters). Data clustering techniques – such as the self-organizing map (SOM), have been applied to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for exactly the same tap position. SOM has been applied to detect abnormal conditions and to hypothesize about the nature of the abnormalities.

- Data mining methods have been applied to dissolved gas analysis (DGA) in power transformers. DGA, as a diagnostics for power transformers, has been available for many years. Methods such as SOM has been applied to analyze generated data and to determine trends which are not obvious to the standard DGA ratio methods (such as Duval Triangle).

- In educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning, and to understand factors influencing university student retention. A similar example of social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized, and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate institutional memory.

- Data mining methods of biomedical data facilitated by domain on tologies, mining clinical trial data, and traffic analysis using SOM.

- In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction

incidents. Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.

• Data mining has been applied software artifacts within the realm of software engineering: Mining Software Repositories.

Human rights

Data mining of government records – particularly records of the justice system (i.e., courts, prisons) – enables the discovery of systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records by various government agencies.

References

1. Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research". Informing Science 3 (2).
2. Foote, Jonathan (1999). "An overview of audio information retrieval". Multimedia Systems (Springer).
3. Beel, Jöran; Gipp, Bela; Stiller, Jan-Olaf (2009). "Information Retrieval On Mind Maps - What Could It Be Good For?". Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09). Washington, DC: IEEE.
4. Frakes, William B. (1992). Information Retrieval Data Structures & Algorithms. Prentice-Hall, Inc. ISBN 0-13-463837-9.
5. Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.
6. Zhu, Mu (2004). "Recall, Precision and Average Precision".
7. Turpin, Andrew; Scholer, Falk (2006). "User performance versus precision measures for simple search tasks". Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, WA, August 06–11, 2006) (New York, NY: ACM): 11–18. doi:10.1145/1148170.1148176. ISBN 1-59593-369-7.
8. Everingham, Mark; Van Gool, Luc; Williams, Christopher K. I.; Winn, John; Zisserman, Andrew (June 2010). "The PASCAL Visual Object Classes (VOC) Challenge".International Journal of Computer Vision (Springer) 88 (2): 303–338. doi:10.1007/s11263-009-0275-4. Retrieved 2011-08-29.
9. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.

10. K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann (2010). The binormal assumption on precision-recall curves. Proceedings of the 20th International Conference on Pattern Recognition, 4263-4266.
11. Mooers, Calvin N.; Theory Digital Handling Non-numerical Information (Zator Technical Bulletin No. 48) 5, cited in "information, n.". OED Online. December 2011. Oxford University Press.
12. Doyle, Lauren; Becker, Joseph (1975). Information Retrieval and Processing. Melville. pp. 410 pp. ISBN 0-471-22151-1.
13. Maron, Melvin E. (2008). "An Historical Note on the Origins of Probabilistic Indexing". Information Processing and Management 44 (2): 971–972. doi:10.1016/j.ipm.2007.02.012.
14. N. Jardine, C.J. van Rijsbergen (December 1971). "The use of hierarchic clustering in information retrieval". Information Storage and Retrieval 7 (5): 217–240. doi:10.1016/0020-0271(71)90051-9.
15. Korfhage, Robert R. (1997). Information Storage and Retrieval. Wiley. pp. 368 pp. ISBN 978-0-471-14338-3.

"The lesson content has been compiled from various sources in public domain including but not limited to the internet for the convenience of the users. The university has no proprietary right on the same."

9, Km Milestone, NH-65, Kaithal - 136027, Haryana
Website: www.niilmuniversity.in