Chapter 6

Application integration in applied text mining

D. Sullivan
Ballston Group, President,
United States.

Abstract

Text mining will be adopted in the enterprise only when it effectively integrates with existing applications. This chapter describes three classes of text mining applications and a common integration model for extracting content, pre-processing text, performing linguistic analysis and supporting end user analysis in all three. Storage management and access control issues are also addressed.

1 Introduction

As text mining expands from primarily academic domains into the realm of applied information technology we need to fit this discipline's techniques into the broader infrastructure of enterprises. Questions about parsing techniques, lexical knowledge bases and other aspects of natural language processing will of course continue to demand our attention but application integration challenges will also be at the forefront of applied text mining projects. In this chapter we will examine how text mining systems fit into the application mix of large enterprises and discuss the basic processes involved in integrating these applications with other enterprise and external systems.

The next section begins with a discussion of different types of text mining applications and the business drivers behind them. This is not an exhaustive taxonomy; applied text mining is too immature for that. In section three we examine the broad level application stages in applied text mining including content acquisition, pre-processing, linguistic analysis, content storage and analysis, as well as security, access control and rights management issues that may arise in some text mining projects.

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/06

2 Business drivers and application types

Information technology (IT) applications exist to solve specific business problems, such as tracking customer accounts, complying with government regulations, monitoring health and environmental safety. For the past decade, organizations have exploited business intelligence (BI) applications, such as ad hoc query tools, data warehouses and on-line analytic processing systems to improve performance and productivity. BI has been widely successful and will continue to expand its reach within organizations. Traditional BI, though, with its emphasis on structured data will never meet all the needs of large scale organizations; there is simply too much variety and complexity in customer relations, for example, to adequately codify all possible complaints in a structured data representation. Notes and other free form text commentary are frequently required to provide details for those course level categorization schemes. As unstructured text, this information is inaccessible to most BI and data mining techniques. Text mining operations, such as feature extraction and information extraction, map the key elements of text to a structured format that lends itself to well developed analytic techniques and therein lies the promise of applied text mining.

Three broad types of text mining applications in commercial enterprises are:

- Customer transaction analysis
- Competitive intelligence
- Research and development support

Each of these uses similar text mining techniques but require different approaches with respect to application integration.

2.1 Customer transaction analysis

In customer transaction analysis, the object is to extract key entities and facts from free form text and map them to a structured or semi-structured output suitable for aggregate analysis, such as ad hoc reporting, OLAP analysis and data mining. For example, an automobile manufacturer might want to analyze customer comments regarding noise in their subcompact models. Do customers frequently refer to speed, weather conditions, number of passengers, time of day? A health care insurer might analyze descriptions of on-the-job injuries to determine if environmental conditions correlate with accidents.

This type of application uses internal data sources that typically contain both structured and unstructured data stored in relational databases and business process specific applications. As in data warehousing, text mining in customer transactions requires structured queries to identify transactions; extraction, transformation and load processes, access to multiple systems and the ability to merge multiple data streams.

The goal of customer transaction analysis is to analyze large volumes of comments closely related to structured data.

2.2 Competitive intelligence

Competitive intelligence (CI) is the study of competitors, partners, markets, research, regulatory agencies and other institutions that can affect a business' market position. Common questions in CI include: what are competitors plans for moving into a particular market? Who is performing research in a particular area and willing to partner? What prior art exists related to a fabrication technique about to be submitted for patent approval? Unlike customer transaction analysis, CI depends primarily on external sources of data such as news feeds, government databases, competitors' publications, and industry analyst reports. It requires several techniques to gather raw, relevant CI, including federated searching, crawling, filtering, and categorization. External sources typically have a higher ratio of non-useful and sometimes contradictory information that text mining practitioners need to manage.

The goal of text mining in CI is to extract a broad range of terms and facts from a variety of sources to provide an overall perspective on the strategic position of organizations in a market.

2.3 Research and development support

The pace of scientific research is rapidly outpacing the ability of scientists to keep up with even narrow domains and the life sciences, pharmaceuticals and bioinformatics seem particularly plagued with an overwhelming growth of knowledge. Text mining applications in research and development tend to focus on basic scientific facts. For example, bioinformatics researchers use text mining to identify macromolecules, processes and key relationships. The following is typical of the text found in medical research abstracts:

> The three proteins that comprise anthrax toxin, edema factor (EF), lethal factor (LF), and protective antigen (PA), assemble at the mammalian cell surface into toxic complexes. After binding to its receptor, PA is proteolytically activated, yielding a carboxyl-terminal 63-kDa fragment (PA(63)) that coordinates assembly of the complexes, promotes their endocytosis, and translocates EF and LF to the cytosol [1].

Ideally, a text mining application would extract relevant entities, *e.g.* anthrax toxin, edema factor, lethal factor and protective antigen; the corresponding appositions EF, LF and PA; and the relationship between the entities, such as activation, promotion and coordination. The information extracted from this abstracted would combine with information extracted from myriads of other sources to produce a database of facts about the subject under study. This structured database in turn would be used to gain insight into larger research questions, such as mapping the metabolic pathways of an organism using techniques such as link analysis and visualization.

Each of these application types have specific requirements but when implemented in a production environment, require similar application

integration. The next section provides an overview of the common steps and integration patterns in applied text mining with some discussion of the particular needs of the three broad types of applied text mining applications.

3 Application elements

The application elements of a text mining system are determined by the business driver behind the project. The first step of course is to identify the business need, such as reducing warranty costs, monitoring customer complaints, filtering email, and managing intellectual property. The driver in turn determines what types of information should be mined, such as term correlations, entities or complete facts. With an understanding of the overall need and the specific types of information sought, specifying the details of the implementation can begin.

The application elements are closely aligned with the basic text mining process steps:

- Content acquisition
- Pre-processing
- Linguistic analysis
- User analysis

and ancillary processes such as

- Content repository management
- Security and access control

Each of these elements, depicted in fig. 1, will be outlined below.

3.1 Content acquisition

The first step in text mining is acquiring content for analysis. Depending upon the requirements, this content can be found in multiple sources, some internal and others external.

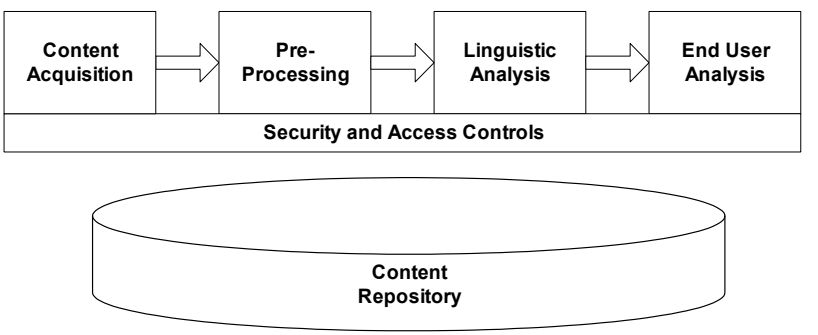

Figure 1: Logical elements of an integrated text mining application.

3.1.1 Internal content acquisition

Internal data are typically found in strategic applications, such as customer relationship management (CRM) and sales force automation (SFA) systems. Gateways, such as ODBC, extract free form text along with structured attributes from these database applications and stage that data in an intermediate storage area for pre-processing. As in data warehousing, one typically starts with a full, initial extraction followed by incremental updates. When planning internal data acquisition functionality, one must consider the time required to complete the extraction process, other requirements on the source application and the sometimes limited window of opportunity for performing the extraction, as well as read access to all relevant attributes. If text is extracted from multiple systems then reference identifiers, such as primary keys, must be extracted as well to allow the merging of multiple data streams.

3.1.2 External content source

Acquiring data from external sources poses a different set of technical and rights management issues than internal sources.

External content sources are typically on-line and accessible via crawling or federated searching. Crawlers download content and follow hyperlinks within a document or page to other content which is in turn downloaded and its links are followed, and so on. These tools are effective for gathering content from small sites, such as a Web site with a conference's proceedings. These tools can, however, put an inordinate demand on Web servers and some sites do not allow their use. (The Robot Exclusion Protocol was developed to specify crawler operations allowed at a site.)

Crawlers copy content and so require storage and indexing services on the host system running the crawler. Federated search avoids these problems.

Federated searching is the process of using multiple search engines to find relevant content. When a user specifies a query to the initiating search engine, that system sends the query to multiple client search engines. Each search engine in turn executes the query and returns the results. The initiating server then reformats, merges and ranks the results before presenting results to the user. One key advantage of federated search is scalability. Content is not downloaded from multiple repositories and additional search engine indexes are not required. The cost of adding additional client search engines is the cost of the gateway between the servers.

With query languages, federated search engines provide greater control for identifying relevant content than crawlers which provide only course parameters, *e.g.* hosts names and number of links to follow. Since search engines typically return links and descriptions of matching documents, crawlers can easily download documents included in result lists.

3.1.3 Rights management

Regardless of which method is used, content is often protected by

copyrights. Some content may include a description of rights using the Dublin Core or other metadata specification. In other cases, a copyright statement is embedded in the text. Regardless of how rights are specified in text (if they are specified at all), care must be taken to abide by the limitations imposed by copyright owners.

3.2 Pre-processing

The pre-processing phase includes all operations that must be performed on text after it is acquired and before it can be analyzed with text mining algorithms. These operations include: removing extraneous content, extracting text from proprietary formats, mapping to an XML format and annotating with derived data and metadata.

On-line content often contains text that is irrelevant from a text mining perspective. Web pages may have embedded Javascript functions, formatting tags, and irrelevant URLs (*e.g.* links to home pages) that should be removed completely. Header information, such as journal names, authors, and publication dates should be mapped to metadata attributes.

Documents in proprietary formats, such as Microsoft Word or Adobe PDF, should be mapped to plain text during pre-processing. A number of commercial products are available to extract content from commonly used file formats and while these work well in most cases, there are some limits. For instance, it is especially challenging to extract text while preserving the semantics of tabular information. Also, images frequently contain labels that are not extracted by reformatting programs and so valuable information is lost.

Mapping text to XML formats is not required to mine text but it does offer several advantages. First, it is easier in later phases of text mining to extract particular sections of text, for example, the abstract of a scientific paper, the XXX section of a patent or the YYYY description from a Securities and Exchange Commission 10-K report. Second, metadata attributes are easily embedded along with the original text. Most relational database applications support importing and exporting to XML schemas so the mechanics of persistent storage are less tedious even when a native XML database is not used.

The final step of pre-processing is annotating with derived data and metadata. Derived data can be as simple as word counts or as complex as categorization labels. Metadata is often available directly from text, such as authors' names, publication dates, and name of publication. Some proprietary file formats include metadata attributes, in the case of HTML or plain text, they have to be extracted from the text using techniques such as regular expression matching. These attributes will later serve to improve text mining operations by allowing finer grained control over the text analysis phase. For example, an analyst might want to select abstracts with particular keywords, written during a specific time period and categorized with a particular set of labels.

At the end of the pre-processing phase, content has been acquired, extraneous text has been removed, content has been mapped to a semi-structured format and derived data and metadata have been explicitly

incorporated into the text. The focus now shifts from applications that perform low data manipulations to those that analyze the text using linguistic processing techniques.

3.3 Linguistic analysis

The details of linguistic analysis in text mining are well documented in other sections of this book and will not be restated here. Instead, this section will discuss basic operations with respect to application integration. If the linguistic analysis component is treated as a black box, then we are left to examine the inputs, which we have just described in content acquisition and pre-processing, and the outputs, which is the primary topic of this section.

The purpose of the linguistic analysis component is to extract terms, entities and facts from texts and structure their representation in such a way as to facilitate analysis.

3.3.1 Term co-occurrence

At the term extraction level, text mining tools typically output weighted pairs of co-occurring terms. The terms are often stemmed and so in a basic canonical form. The weights are a function of the frequency with which terms occur together and the proximity of the terms in the text. Term co-occurrence measures are useful when combined with visualization techniques to explore broad, unfamiliar domains.

3.3.2 Entity extraction

Entity extraction identifies multi-term names of persons, places, organizations, dates, monetary amounts and other objects. Tools frequently provide a measure of the frequency of which the entities appeared in the text, a canonical form for entity (*e.g.* the canonical form of U.N. is United Nations) and the type of entity. While more useful than term co-occurrence because it operates at a semantic level rather than a lexical level, entity extraction does not provide information about the relationship between the entities. Consider the phrase from the previously discussed scientific abstract,

> …PA is proteolytically activated, yielding a carboxyl-terminal 63-kDa fragment…

Entity extraction would extract ‘PA’, probably in canonical form ‘protective antigen’, and ‘carboxyl-terminal 63-kDa fragment’ but without indicating how they relate. Information extraction addresses that issue.

3.3.3 Information extraction

Information extraction targets patterns of entities and their relationships, such as companies engaging in mergers and enzymes initiating metabolic processes. Information extraction builds on entity extraction by analysing relationships

between entities as described by the predicate in sentences. In the above example, 'activated' is a key term that indicates an agent is initiating a process in a target object. The output of these systems are often XML structures such as:

```
<activation>
  <agent> protective antigen </agent>
  <object>carboxyl-terminal 63-kDa </object>
  <mode> proteolytically</mode>
<activation>
```

This fact in turn can be combined with other facts, such as how carboxyl-terminal-63-kDa affects another entity which in turn affects another and so on.

This type of analysis is addressed next.

3.4 User analysis

The last applications of an integrated text mining environment are user analysis systems. There are three broad types: ad hoc query tools, link analysis tools, visualization tools. Data mining tools that use structured attributes extracted from text would also fall into this category but will not be discussed further.

Ad hoc query tools are commonly used in data warehouse reporting systems and can easily extend to meet the needs of some text mining applications. These tools are particularly useful when terms frequencies and entities are extracted from free form text in a structured record, such as in CRM or insurance claim applications.

Link analysis tools are appropriate for identifying paths of relationships between entities. For example, an information extraction tool working with medical abstracts may find that A activates B, B promotes C, and C induces D. Link analysis tools allow analysts to explore the complex chains of interaction among entities.

Visualization tools are especially useful for analyzing term co-occurrences. Terms that occur frequently together are in close proximity in 2-dimensional spaces. The relative importance of a topic or group of terms can be reflected using a 3rd dimension.

The choice of end user analysis tools reflects the variations and distinctions in text mining techniques and business objectives that drive their adoption.

3.5 Content repository

Behind the process of content acquisition, pre-processing, linguistic analysis and user analysis must reside a content repository. The structure of this repository will depend on a number of factors.

First, will the analyzed content be stored along with extracted and derived information? There may be no reason to store publicly available patents or scientific abstracts in full, a link to the source is sufficient. If there is a chance

the original content could be lost, for example if customer records are deleted or archived after an account is closed, then the text should be maintained within the text mining application.

Second, is the extracted information supplementing an existing database, such as a CRM record, or is it part of an independent application, such as an R&D support system? In the latter case an independent XML or relational database would be required for the system.

Third, what are the access control requirements? Information derived by text mining will not necessarily have the same security requirements as the source systems of the original text.

Storage management, like security and access controls, affect each component in an integrated text mining environment.

3.6 Security and access controls

Security issues should be addressed in the early stages of system design. Clearly, processes that gather content require read access to repositories but one will have to decide whether a single account will be granted read access to content across all sources of text or will multiple account share privileged access. The latter option reduces the cost of a security breach on a single account but requires marginally more administration.

Access controls should also account for rights management requirements. Crawling and harvesting processes should be run in accordance with contractual agreements with on-line subscription services.

Information derived through text mining processes should be managed according to access requirements. For example, clustering and term co-occurrence information may be broadly available for general use with visualization tools as an aid to search and navigation. These data should be managed separately from derived facts used to drive competitive intelligence analysis or other strategically sensitive applications.

Security issues are pervasive in enterprise systems and text mining applications must operate within the frameworks dictated by those issues.

4 Conclusions

Applied text mining systems function in the broad matrix of enterprise applications where integration is essential. There is a broad range of text mining applications including customer transaction analysis, competitive intelligence and research and development support and they can all be addressed with a similar system integration model. This chapter has described a basic design pattern modelled after the architectural principals developed in data warehousing and enterprise search. The pattern includes content acquisition, pre-processing, linguistic analysis and end user analysis with storage management and access control mechanisms spanning these components.

References

[1] Mogridge, J., Cunningham, K., Lacy, D.B., Mourez, M. & Collier, R.J., *Proceedings of the National Academy of Science* U.S.A. 2002 May 14;99(10):7045-8.