www.db2mag.com
DB2 magazine
DB2 is a registered trademark of IBM and is used here under License

Quarter 2, 2002

http://www.db2mag.com/db_area/archives/2002/q2/miner.shtml

Words of a Feather

Applying text and data mining to the terms we use to describe things can reveal which birds (or companies or people) will flock together.

By Charles Zhou and Ying Zhao

People use many words to describe an object. The words can be people's names or company names. They can describe shape, size, or any other characteristic. And human intelligence can digest this information and extract relationships among objects based on the words that describe key characteristics. That human intelligence works perfectly well if the input information is limited.

Resources

Intelligent Miner for Text
ibm.com/software/data/iminer/fortext

Intelligent Miner for Data
ibm.com/software/data/iminer/fordata

But we all know that information sources are anything but limited. If you happen to be looking for financial news related to your stock portfolio, you'd find more than 2,000 sources on the Internet. No single person can digest all that information every day.

In fact, trying to follow the financial news that might affect our portfolios started us thinking about how to train a computer to automate this human ability and process far more information than our minds can. Our exploration led us to build a process of text mining integrated with data mining that extracts features and relationships among objects based on their text descriptions. We called this process *knowledge space intelligence* (taking the term "knowledge space" from the theory introduced by Jean-Paul Doignon in 1985). Figure 1 illustrates how the process works: crawling Internet sites, collecting information, extracting features, building a knowledge space, mining the knowledge space, and visualizing results.

Figure 1

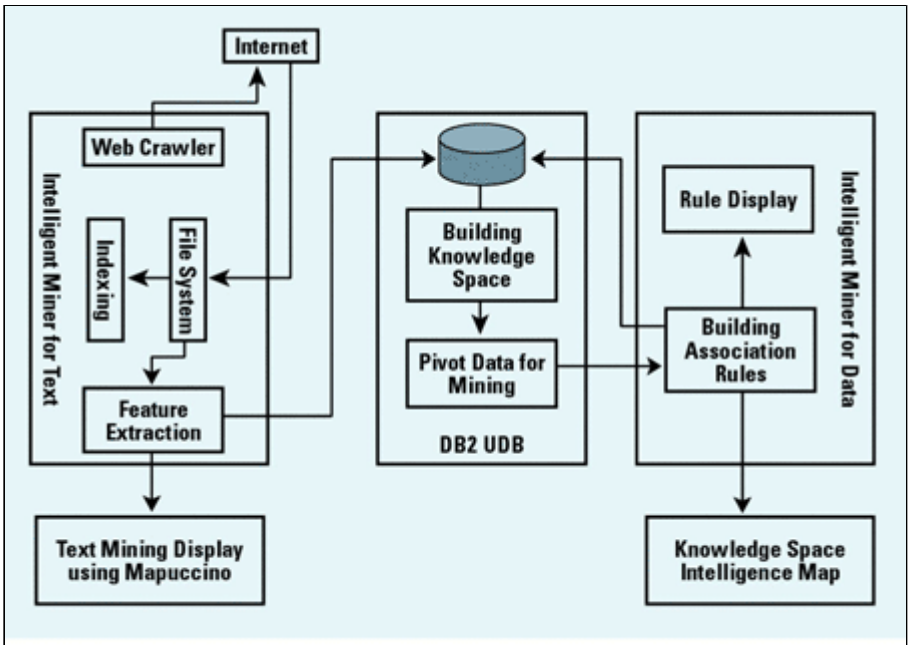


Figure 1: A knowledge space intelligence architecture.

KNOWLEDGE SPACE INTELLIGENCE

Let's take an example to illustrate how knowledge space intelligence works. For objects A, B, and C, we use the words red, square, and heavy to describe their color, shape, and weight. A set of words is used to describe each object: {Square} for object A; {Square, Heavy} for object B; and {Red, Heavy} for object C. The set is the knowledge space for that object. To discover the relationship between two objects (A and B), you match the words that describe both A and B. In this example, 100 percent of the words that describe A also describe B (Square). Similarly, 50 percent of the words that describe B also describe C (Heavy). No words that describe C also describe A. The strength of an association can be measured by the percentage of overlapping words in the knowledge spaces. The similarity among objects is equivalent to the similarity among their knowledge spaces.

Knowledge space intelligence is a process that takes the knowledge spaces as input and derives association rules among the objects. Based on the strength of each association rule, a filter is set. The set and the association rules can be represented in a knowledge space intelligence map.

CONSTRUCTING A KNOWLEDGE SPACE

You can build the knowledge space for an object by collecting information about the object from many texts. If the collected information for the object is already in a structured format such as a database table, knowledge spaces are simply the records in the database tables.

In the real world, however, most of our knowledge about objects around us comes from text descriptions of these objects. Most texts aren't in structured database formats. Therefore, we need to extract a knowledge space from the text descriptions using feature extraction techniques. Features are key elements of a text that represent some meaningful entity. Feature extraction takes a text description of an object as input and automatically extracts key words, key terms, people, places, and organizations that characterize the object.

To demonstrate this process, we used a module called Text Analysis Tools built into Intelligent Miner for Text. The text analysis tools collect features from a document or a collection of documents, including key words, key terms, people, places, organization names, time, date, currency, relations, and so on. The tool normalizes the features it finds and groups together occurrences of features if they refer to the same entity or concepts. For example, the feature extraction tool finds "Bill Clinton" and "William Clinton" in documents. It then maps "Bill Clinton" to "William Clinton" in the knowledge space, where "Bill Clinton" is treated as an infected form of "William Clinton." A feature extraction tool must have the capability to map the infected form of words to their canonical forms. Feature extraction is the foundation for knowledge space intelligence.

DISCOVERING ASSOCIATIONS

Once you've created the knowledge spaces for a set of objects, you can mine them. We used an association algorithm in Intelligent Miner for Data to discover association rules for the objects in their knowledge space. The input to the association algorithm is a table with features and objects (see Table 1).

Table 1 *Input to an association algorithm.*

Features (Transaction ID)	Objects (Items)
Red	C
Square	A
Square	B
Heavy	B
Heavy	C

The typical application of association rules is to discover product associations and affinities from business transaction data. There are usually two columns of input data to the algorithm: *transaction ID* and *items*. Transaction ID represents the unit to group the items, such as a market basket or an order. Items represents products that are bought in a transaction. In knowledge space mining, substitute *features* for transaction ID and *objects* for items. Table 2 shows the association rules of objects A, B, and C.

Table 2 *Association rules.*

Body	Head	Support %	Confidence %	Expected Confidence of Head %	Lift
A	→ B	33 {1 Square}	100	67	1.50
B	→ A	33 {1 Square}	50	33	1.50
B	→ C	33 {1 Heavy}	50	67	0.75
C	→ B	33 {1 Heavy}	50	67	0.75

For each pair of associations, there are four numbers that measure the strength of the association: support, confidence, expected confidence, and lift. Support measures the percentage of features that describe both objects. Confidence is a concept of conditional probability for the occurrence of one object (Head, in Table 2) while the other (Body) occurs. It is a percentage of features that describe the object in the left hand side of a rule that will also describe the object in the right hand side of the rules. Lift is the ratio of the confidence of a rule over the expected confidence, which is the overall percentage of features that describe the object in the right hand side of the rule. The strength of object A associated with object B is usually measured with all three metrics. However, for simplicity, the knowledge space intelligence map we will show you uses only confidence to measure the strength of an association.

We express the expected confidence of A as *P(A)*, and the expected confidence of B as *P(B)*, where *P(A)* and *P(B)* are probability functions of objects A and B in the knowledge space. Then:

Lift(A → B) = (confidence of B) / (expected confidence of B)
= P(B/A) / P(B)
= P(A ⊗ B) / {P(A) P(B)}
where *P(A ⊗ B)* is called support for objects A and B.

The knowledge space intelligence can be represented in the Associations Visualizer in Intelligent Miner for Data. In order to better understand the association rules in the knowledge space, and to visualize the "if body, then head" relation, we developed a Java applet to create a knowledge space intelligence map that visually displays the association rules discovered from a knowledge space.

A KNOWLEDGE SPACE EXAMPLE

Let's apply these concepts to some textual information. In this simple example, we'll extract seven features from three business articles: database, chips, Unix, consulting, PC, network, and business intelligence. (Of course, we aren't providing advice about buying or selling any security. We're simply demonstrating an aid for digesting massive information quickly.)

In a complete knowledge space, the features may number several hundred or thousand terms. Table 3 shows the knowledge space for IBM, Oracle, Microsoft,

Intel, and Sun Microsystems. Using this table as input, you can calculate the confidence, support, expected confidence, and lift, using the features and companies (objects) as input.

Table 3 *The knowledge spaces of five companies extracted from business articles.*

Features	Firms Mentioned in the News (Objects)				
	IBM	Oracle	Microsoft	Intel	Sun
Database	x	x	x		
Chips	x			x	x
Unix	x				x
Consulting	x	x	x		x
PC	x		x	x	
Network	x		x	x	
BI	x	x	x		

Table 4 shows the resulting association rules. As an example, let's look at the association rule of Sun→Microsoft on line three of Table 4. Of the seven features we've selected, Sun has three and Microsoft has five. One out of three features that describe Sun also describes Microsoft (Consulting). Therefore, the confidence of Microsoft in this rule is P(Microsoft/Sun) = 1/3. The expected confidence of Microsoft is P(Microsoft) = 5/7. The lift of this rule is the ratio of confidence over expected confidence: Lift (Sun→Microsoft) = (1/3)/(5/7) = 7/15, which indicates how similar the two companies are. If the lift is greater than 1, the two companies are very similar; if the lift is less than 1, the two companies aren't very similar. The information from these three articles doesn't show much similarity between the two companies.

Table 4 *Association rules derived from Table 3.*

Body	Head	Confidence	Support	Expected Confidence	Lift
Sun	→ Intel	1/3	1/7	3/7	7/9
Intel	→ Sun	1/3	1/7	3/7	7/9
Sun	→ Microsoft	1/3	1/7	5/7	7/15
Intel	→ Microsoft	2/3	2/7	5/7	14/15
Microsoft	→ Intel	—	—	—	—
Intel	→ IBM	3/3	3/7	7/7	1
IBM	→ Oracle, Sun	1/7	1/7	1/7	1
Oracle	→ Sun	1/3	1/7	3/7	7/9
Intel, Microsoft	→ IBM	2/2	2/7	7/7	1

When these association rules are constructed from thousands of articles, some distinctive relations will show up between one company and a few others among thousands of companies.

THE BUSINESS APPLICATION

To show how knowledge space intelligence might work in the real world, we built a demo system called IntelliNews that collects all the business news from a Web site (such as Yahoo) and discovers associations among companies from Internet news. The goal of this application case is to examine relations between two or more of the companies mentioned in the recent business news. To do so, IntelliNews carries out the following five steps:

1. Collects recent business news items from Yahoo using a Web crawler in Intelligent Miner for Text.

2. Extracts companies from the news (in this case, we extract only public companies with stock ticker symbols — between 500 and 2,000 companies appear in Yahoo business news every day).

3. Extracts knowledge spaces for the companies in the news using the feature extraction tool in Intelligent Miner for Text. Each company is treated as an object. The feature extraction tool extracts names, organizations, places, and relations that appear in the business news. Figure 2 shows a high-level view of the knowledge space for I2 Technologies (identified by its stock ticker ITWO). Figure 3 shows details of some of the categories in the I2 knowledge space.

Figure 2: *The visual knowledge space for I2 Technologies (ITWO).*

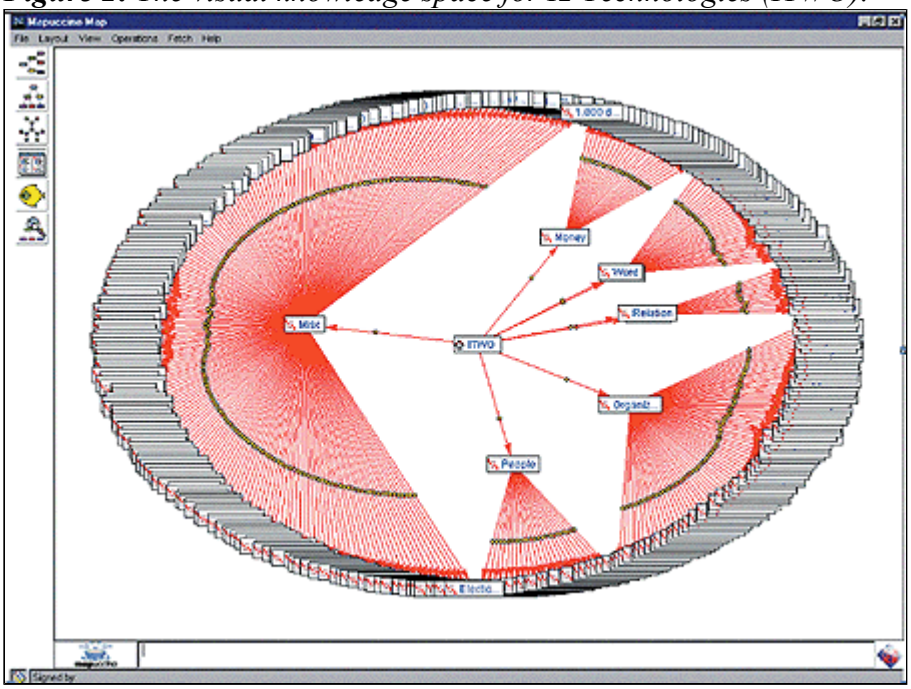
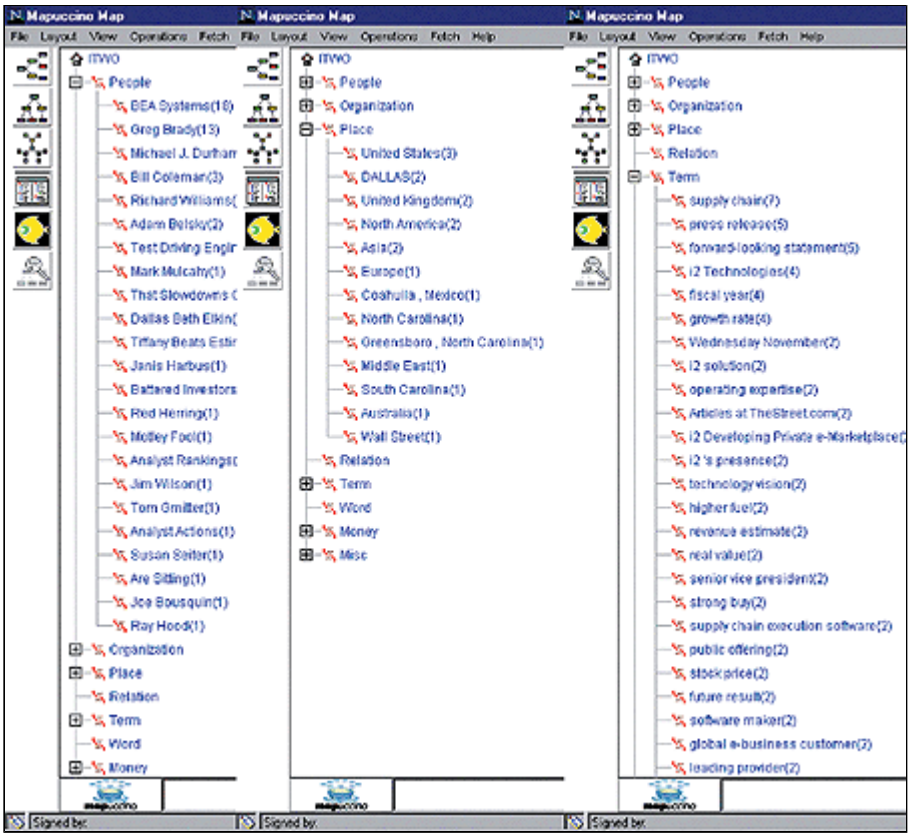


Figure 3: *Knowledge space details for I2 Technologies (ITWO).*

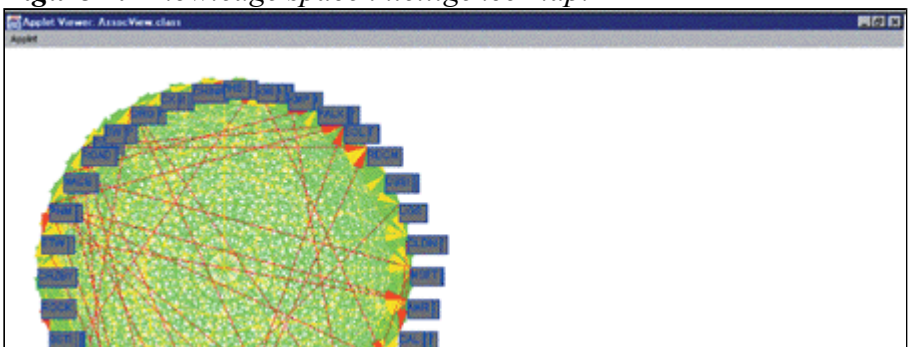


4. Discovers association rules for the companies from the knowledge spaces.

5. Visualizes the knowledge space intelligence using a Java applet.

Association rules that can also be shown in a table can also be shown as an interactive knowledge space intelligence map that makes the associations easier to grasp. In Figure 4, all the associations start in a circle. Each node represents a company. Links represent the associations between two companies. The strength of association is measured using confidence as numbers in the link by clicking the Association check box. The check box Out implies all the associations from one node to other nodes (Fan Out). Similarly, In means all the associations from other nodes to the focal node (Fan In).

Figure 4: *Knowledge space intelligence map.*



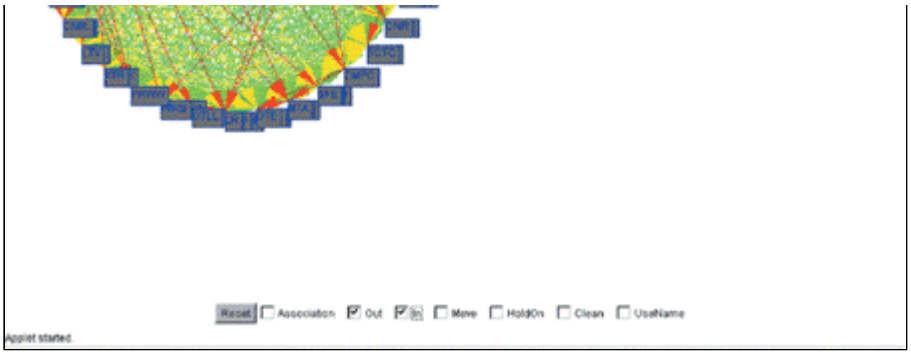
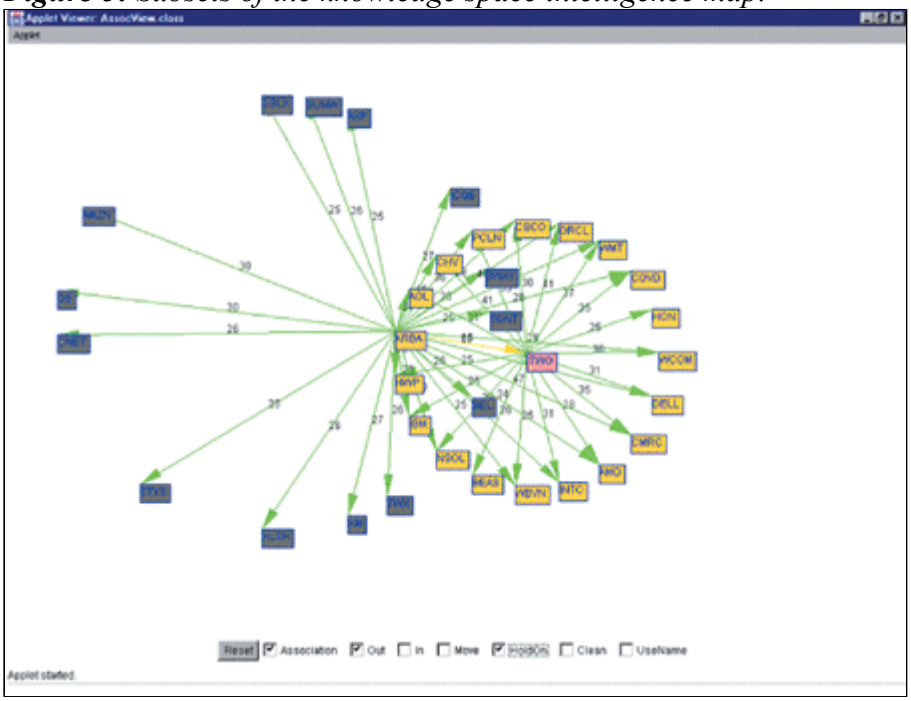


Figure 5 shows subsets of the associations, starting with node ARBA with only the Out box checked. The numbers on links represent the confidence of association rules. All the companies that associated with ARBA are circled around ARBA. One of them is ITWO. If you check Hold On and click the node ITWO, the associated companies for ITWO are circled around ITWO. The associations are very consistent with the reality of the business world, which shows how you can use the functions to track a chain of associated companies.

Figure 5: *Subsets of the knowledge space intelligence map.*



ENDLESS POSSIBILITIES

The integrated knowledge space intelligence framework we demonstrated combines text and data mining technology for an interesting application: to discover associations among companies from recent Internet business news. Such associations can give savvy investors early hints of possible mergers between companies or give businesses an idea of activity among their rivals. In fact, when

you consider how much text is around to analyze, the possibilities are limited only by your imagination.

Charles Zhou, Sc.D., is a consultant in DB2 and business intelligence for IBM's customers and Data Management partners. You can reach him at czhou@us.ibm.com.

Ying Zhao, Ph.D., is currently with Quantum Intelligence Inc. You can reach her at ying_zhao@quantumintelligence.com.

Return to Article