



The best of both worlds: Probabilistic and rules-based text mining

4th September 2002



Automated approaches to text mining

Text mining encompasses both:

- Search technologies
- Classification technologies

Each of these technologies can be probabilistic, rules-based or both



Rules-based searching

- **Search processes are often unnatural**

eg. Boolean searches can be complex and unwieldy - even for information professionals

- **Volume**

Search engines will list a document as a match even if each of the words matching the query only occur once in the document. The resulting list can be 1000s of documents long

- **Search words may be synonymous with, but not exact matches to corresponding words used in text**

eg. If you search under 'red', documents containing 'scarlet' or 'russet' will not be returned



Probabilistic searching - Smartlogik Discover

- Free text searching (with Boolean key word features)
Understands the relative importance of words in a given query.
- Concept extraction during indexing enables phrase searching
Essence of each document distilled using:
 - word stemming
 - stop word removal
 - word position analysis algorithms



Probabilistic searching - Smartlogik Discover

- Dynamic word weighting

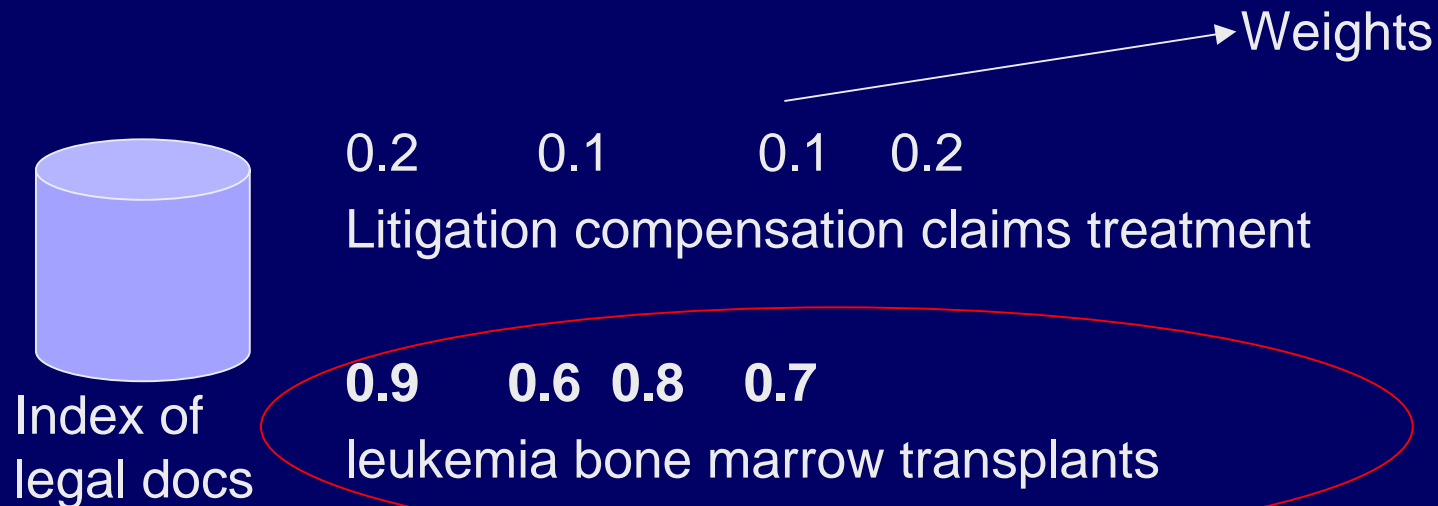
Permits searches to discriminate between terms of greater and lesser relevance within a query.



Dynamic word weighting example

Search for:

Litigation and compensation claims following the treatment of leukemia and bone marrow transplants



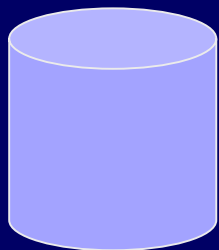
Most important terms



Dynamic word weighting example

Search for:

Litigation and compensation claims following the treatment of leukemia and bone marrow transplants



Index of
medical
docs

0.9	0.8	0.5	0.1
Litigation	compensation	claims	treatment
0.6	0.2	0.4	0.5
leukemia	bone marrow	transplants	

Most important terms



Probabilistic searching - Smartlogik Discover

- Guided search

Relevant terms and categories can be automatically suggested, based on the content of the documents that match the user's search criteria. Users can refine their search quickly and home in on the documents they require, without having to repeatedly modify their queries manually.



Probabilistic searching - Smartlogik Discover

Concepts

Christopher Columbus
discovers America

~~Holidays discovering
the American West~~

~~'Discovery'
credit card~~

(Human)

(Software)

Suggested terms

Columbus

holiday Vegas

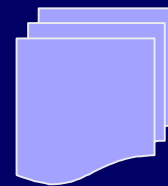
finance

history

~~canyon~~

~~credit~~

Search results



Probabilistic searching - Smartlogik Discover

- 'More like this'

Users can choose specific documents from their search results that are most relevant to their enquiry, and ask to see other similar documents. This process is iterative and allows the user to refine their search indefinitely.



Approaches to classification

- Human
- Probabilistic
- Rules-based
- Combined rules-based and probabilistic



Human approaches

Advantages

- Accurate (for any single classifier)

Disadvantages

- Inconsistent between classifiers
- Expensive
- Time consuming



Probabilistic (fully automated) approach to classification

- Highly automated
- Very low maintenance
- Designed for less complex document and classification set



Probabilistic approach to classification

- Bayesian algorithm allows software to learn how an expert would classify content into a pre-defined structure
- Uses this 'training set' to identify links and correlations between multiple significant words across the document collection
—————▶ probabilistic model of concepts
- Classifies new content in the same way as an expert would



Rules-based approaches to classification

- Designed for complex document set and classification structure
- Transparent, editable classification structure
For example, the structure of the classification system can be made to reflect the organisation's departmental structure in its map of concepts – allowing information to be classified according to the departments that it is relevant to.
- Analyses each document to see if it contains words or phrases in sufficient variety/density to justify classifying it under a particular subject



Rules-based approaches to classification

How are the rules ('rulebases') created?

Initially created by skilled knowledge specialists working with customers, the weighting that individual words carry in deciding how a document should be classified is moderated by their:

- **frequency** – the more often a particular word appears in a document, the more likely that document is to be about that word
- **position** – the higher up a word appears in a document, the more likely that document is to be about that word
- **context** – the relative position of words to one another affect their meaning. For example, the word 'stock' has a number of meanings but if it is preceded by the word, 'beef', it is clear that its context is food-related.



Rulebases



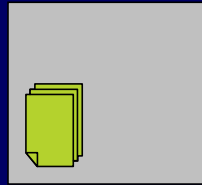
APR's Rulebuilder

- Used to develop rulebases
- Automatic rulebase generation
 - Unlock existing investment in classification systems
 - Quickly create a set of rulebases
- Rule editor
 - Interactively build and test rulebases
 - Refine automatically generated rulebases

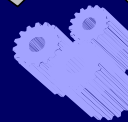
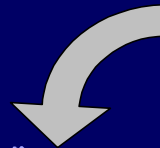
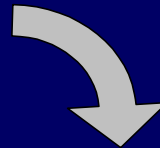
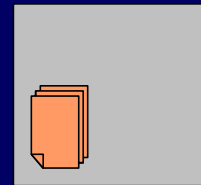


Automatic rulebase generation

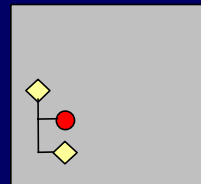
'Positive' set



'Control' set



Comparison



Rulebase



Combined approaches - Smartlogik Structure

- Offer outstanding **precision *and* recall**, dramatically enhancing categorisation and search results.
- Documents can be classified into taxonomies, thesauri, ontologies or vector-based information space
- Automated, configurable classification offers close fit to existing business processes
- Spidering technology enables classification of documents from multiple data feeds, intranet documents and web content



Case study: UK broadcaster

- One of the largest and most diverse news archives in the world
- Used to classify documents manually
- Journalists and programme makers used to have to wait for up to 4 days for the information they requested
- Each document could only be assigned to up to 3 categories because that was the number of copies of the article that were bought



Case study: UK broadcaster

Business objectives

- Robust, reliable, scalable online news archive
- Capable of handling at least 5,000 users and more than 2 million queries per annum
- Retention of existing taxonomy with flexibility to update structure
- Varied user base



Case study: UK broadcaster

- Incorporates both Structure and Discover for unparalleled search accuracy, intelligent refining options and personalised alerts
- Content delivered by LexisNexis
- 16½ million news titles, rising by 6,000 articles per day



Case study: UK broadcaster

SEARCH ⓘ

SUBJECT ⓘ **BROWSE**

HEADLINE

FREE TEXT

WAR ON IRAQ

USE BOOLEAN SEARCH

SCOPE ⓘ **BROWSE**

"POLITICAL AND PUBLIC AFFAIRS"

DATE

FROM

TO (dd/mm/yyyy)

PUBLICATION ⓘ **SOURCE LIST**

"DAILY TELEGRAPH" OR "FINANCIAL TIM

COMPANY ⓘ **BROWSE**

COUNTRY ⓘ **BROWSE**

USA

BY-LINE

CUT OFF %

SEARCH ⓘ

SORT BY

- A journalist carries out a free text search for “War on Iraq’
- He narrows his search by choosing ‘Political and Public Affairs as the scope of his search, and only looking for articles concerning the country, ‘USA’.
- He chooses to search for articles in the national broadsheets dated 7th - 14th August



Case study: UK broadcaster

Documents 1-71 out of 71

PRINT 

saddam. hussein. iraqi. militari.

ADD TOP TERMS 

CUT OFF %

SORT BY

SEARCH 

MORE LIKE THIS 

ALL CLEAR

GET ARTICLES

GET CONTEXT

VIEW 

- | | | |
|---------|-------------------------------------|---|
| 1: 100% | <input type="checkbox"/> | AFGHANISTAN IS ON THE BRINK OF ANOTHER DISASTER Independent - 14/08/2002 (1270 words) |
| 2: 100% | <input checked="" type="checkbox"/> | Business & Media: The business of invading Iraq: War: who is it good for?: Bush is gambling that victory over Saddam will lift the US economy out of double-dip recession - but he risks sparking another oil crisis. Faisal Islam reports Observer - 11/08/2002 (1231 words) |
| 3: 100% | <input checked="" type="checkbox"/> | America's right to fight Iraq: By breaching the terms of the 1991 UN ceasefire, Saddam Hussein has given the US legitimacy under international law to attack, says John Chipman Financial Times - 13/08/2002 (1042 words) |
| 4: 100% | <input type="checkbox"/> | Muslim radicals in Britain issue 'holy war' warning Daily Telegraph - 14/08/2002 (581 words) |
| 5: 100% | <input type="checkbox"/> | Both US parties back away from Iraq war: Top politicians say Bush must prove Saddam is real threat Guardian - 13/08/2002 (685 words) |
| 6: 100% | <input type="checkbox"/> | Mudhafar Amin: Why not put our offer to the test?: The US cannot make war on Iraq without British diplomatic cover Guardian - 08/08/2002 (816 words) |
| 7: 100% | <input checked="" type="checkbox"/> | Comment & Analysis: Why not put our offer to the test?: The US cannot make war on Iraq without British diplomatic cover Guardian - 08/08/2002 (816 words) |


The journalist highlights those documents which most closely match his enquiry and clicks 'More like this', which returns 12 highly relevant documents




Case study: UK broadcaster

Documents 1-12 out of 12 PRINT 

forc. bush. america. american. washington. weapon. state. ADD TOP TERMS 

CUT OFF % SORT BY SEARCH 

MORE LIKE THIS  ALL CLEAR GET ARTICLES GET CONTEXT VIEW 

1: 100%	<input checked="" type="checkbox"/>	Analysis: Iraq conflict: Doves launch last-ditch campaign for Gulf peace: The hawks in Washington have the President's ear - in Europe, calmer voices are speaking out. Jason Burke , Gaby Hinsliff and Ed Vulliamy in New York ask which side Tony Blair plans to back. Observer - 11/08/2002 (2508 words)
2: 100%	<input type="checkbox"/>	Iraq deploys rhetoric to rally support SADDAM HUSSEIN SPEECH DEFIANT LEADER CALLS FOR DIALOGUE WITH UN BUT GIVES NO HINT OF STANCE ON WEAPONS INSPECTORS. Financial Times - 09/08/2002 (530 words)
3: 100%	<input type="checkbox"/>	Both US parties back away from Iraq war: Top politicians say Bush must prove Saddam is real threat Guardian - 13/08/2002 (685 words)
4: 100%	<input type="checkbox"/>	Leading article: No easy choices in Iraq: Saddam is likely to string the US along. Guardian - 13/08/2002 (616 words)
5: 100%	<input type="checkbox"/>	Saudis will not aid US war effort. Guardian - 08/08/2002 (609 words)
6: 100%	<input checked="" type="checkbox"/>	Saddam gives us no choice. Sunday Telegraph - 11/08/2002 (831 words)
7: 100%	<input type="checkbox"/>	Blair not wobbling on Iraq, says aide. Daily Telegraph - 10/08/2002 (615 words)
8: 100%	<input type="checkbox"/>	Iraq crisis : Comment: Saddam diversion spares Bush's blushes. Guardian - 08/08/2002 (723 words)

Finally, he saves his search and sets up a corresponding daily email alert that delivers any new content that matches his search criteria, automatically.



Case study: UK broadcaster

Return on Investment

- 5000+ users p.a. supported on a 24/7 basis
- Direct savings of an estimated £2m p.a. as a result of the removal of the manual news cutting service
- Productivity savings of an estimated £5m p.a., based on 2m searches, 10 minutes/search and a conservative personnel cost of £25k p.a. per researcher
- Handling of in excess of 2m research queries p.a. with the delivery of better quality results through the combined broadcaster/Smartlogik taxonomy



Case study: UK broadcaster

Return on Investment (cont.)

- Significant competitive advantage resulting from the accelerated delivery of relevant information - delivery in seconds, not hours
- Removal of manual archive storage problems and associated risks. Substantially increased archival capabilities for a once limited service
- Flexible delivery and sustainable high levels of service across a very varied user base
- Categorisation, archiving and search consistency to a once imperfect service
- News taxonomy - now a potentially saleable commodity





The best of both worlds: Probabilistic and rules-based text mining

4th September 2002

