CSE 634 – Data Mining:

Text Mining

Group # 8
Presenters:
Munyaradzi Chiwara
Mahmoud Al-Ayyoub
Mohammad Sajjad Hossain
Rajan Gupta

Professor Anita Wasilewska

Outline of Presentation

- **Subject Presentation**
  - **Munyaradzi Chiwara.**
  - **Mahmoud Al-Ayyoub.**
- **Paper Presentation**
  Linguistic Profiling for Authorship Recognition and Verification
  - **Mohammad Sajjad Hossain.**
- **Application Presentation**
  ARROWSMITH
  - **Rajan Gupta.**

References

- Fan, W., Wallace, L., Rich, S. and Zhang, Z. *Tapping into the Power of Text Mining.* Communications of ACM, 2005.
- Grobelnik, M. and Mladenic, D. Text-Mining Tutorial. In the Proceeding of Learning Methods for Text Understanding and Mining, Grenoble, France, January 26 – 29, 2004.
- Hearst, M. *Untangling Text Data Mining.* In the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
- Ramadan, N. M. Halvorson, H., Vandelinde, A. and Levine, S. R. *Low brain magnesium in migraine.* Headache, 29(7):416-419, 1989.
- Swanson, D. R. *Complementary structures in disjoint science literatures.* In the Proceedings of the 14th Annual International ACM/SIGIR Conference, pages 280-289, 1991.
- Witten, I. H. "Text mining." In *Practical handbook of internet computing*, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida, 2005.

References

- Callan, J. A Course on Text Data Mining. Carnegie Mellon University, 2004. http://hartford.lti.cs.cmu.edu/classes/95-779/
- Even-Zohar, Y. Introduction to Text Mining. Supercomputing, 2002. http://algdocs.ncsa.uiuc.edu/PR-20021008-1.ppt
- Hearst, M. *Text Mining Tools: Instruments for Scientific Discovery.* IMA Text Mining Workshop, 2000. http://www.ima.umn.edu/talks/workshops/4-17-18.2000/hearst/hearst.pdf
- Hearst, M. *What Is Text Mining?* 2003. http://www.sims.berkeley.edu/~hearst/text-mining.html
- Hidalgo, J. Tutorial on Text Mining and Internet Content filtering. ECML/PKDD, 2002. http://www.esi.uem.es/~jmgomez/tutorials/ecmlpkdd02/slides.pdf
- Witte, R. *Prelude Overview: Introduction to Text Mining Tutorial*. EDBT, 2006. http://www.edbt2006.de/edbt-share/IntroductionToTextMining.pdf

What is Text Mining?

- **The discovery by computer of new, *previously unknown* information, by automatically extracting information from a usually large amount of different *unstructured* textual resources.**

- **What does *previously unknown* mean?**
  - **Implies discovering genuinely new information.**
  - **Hearst's analogy: Discovering new knowledge vs. merely finding patterns is like the difference between a detective following clues to find the criminal vs. analysts looking at crime statistics to assess overall trends in car theft.**
- **What about *unstructured*?**
  - **Free naturally occurring text.**
  - **As opposed to HTML,XML, …**

Text Mining vs.

- **Data Mining**
  - **In Text Mining, patterns are extracted from natural language text rather than databases.**
- **Web Mining**
  - **In Text Mining, the input is free unstructured text, whilst web sources are structured.**
- **Information Retrieval (Information Access)**
  - **No genuinely new information is found.**
  - **The desired information merely coexists with other valid pieces of information.**

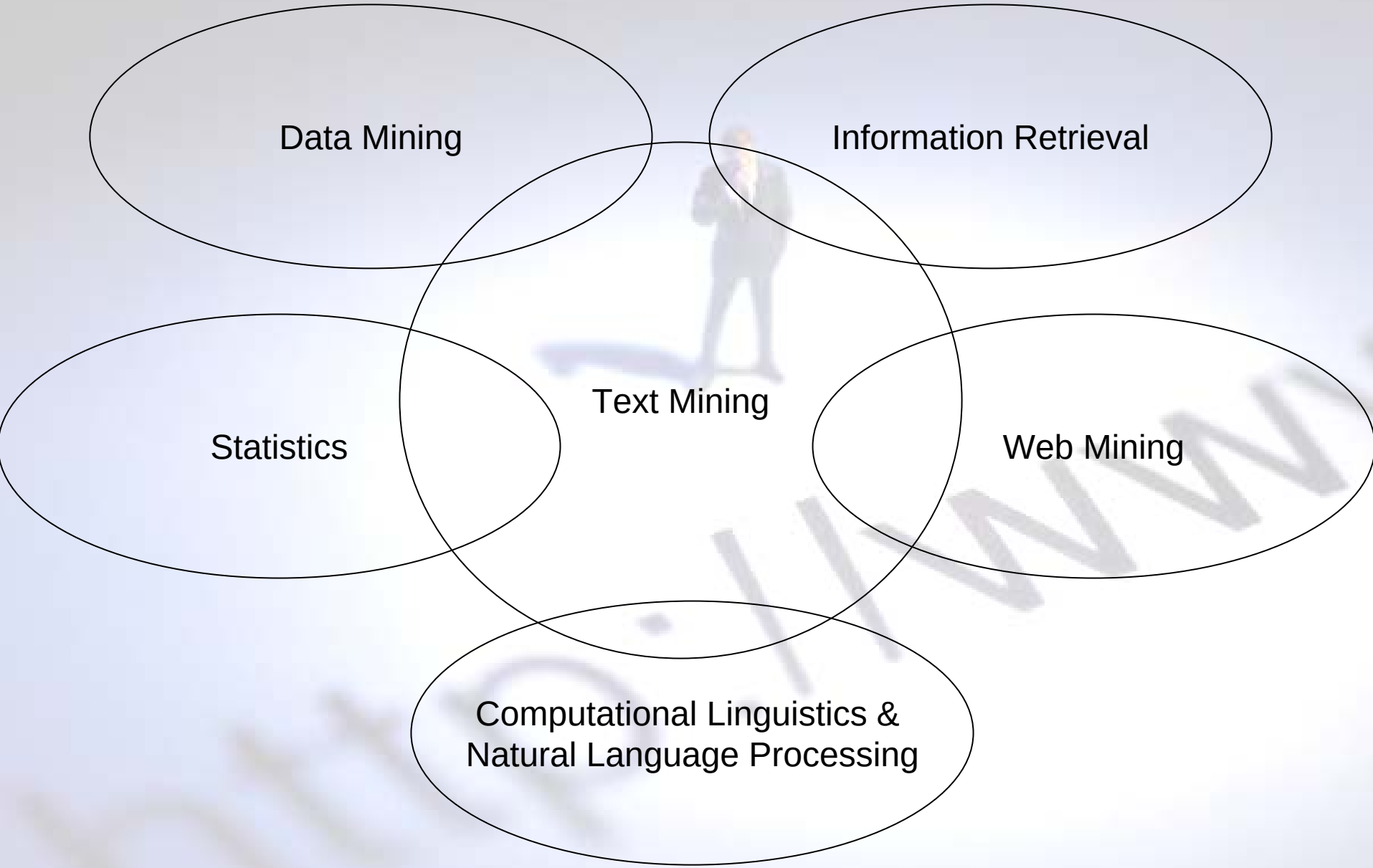
Text Mining vs.

- **Computation Linguistics (CPL) & Natural Language Processing (NLP)**
  - **An extrapolation from Data Mining on numerical data to Data Mining from textual collections [Hearst 1999].**
  - **CPL computes statistics over large text collections in order to discover useful patterns which are used to inform algorithms for various sub-problems within NLP, e.g. Parts Of Speech tagging, and Word Sense Disambiguation [Armstrong 1994].**

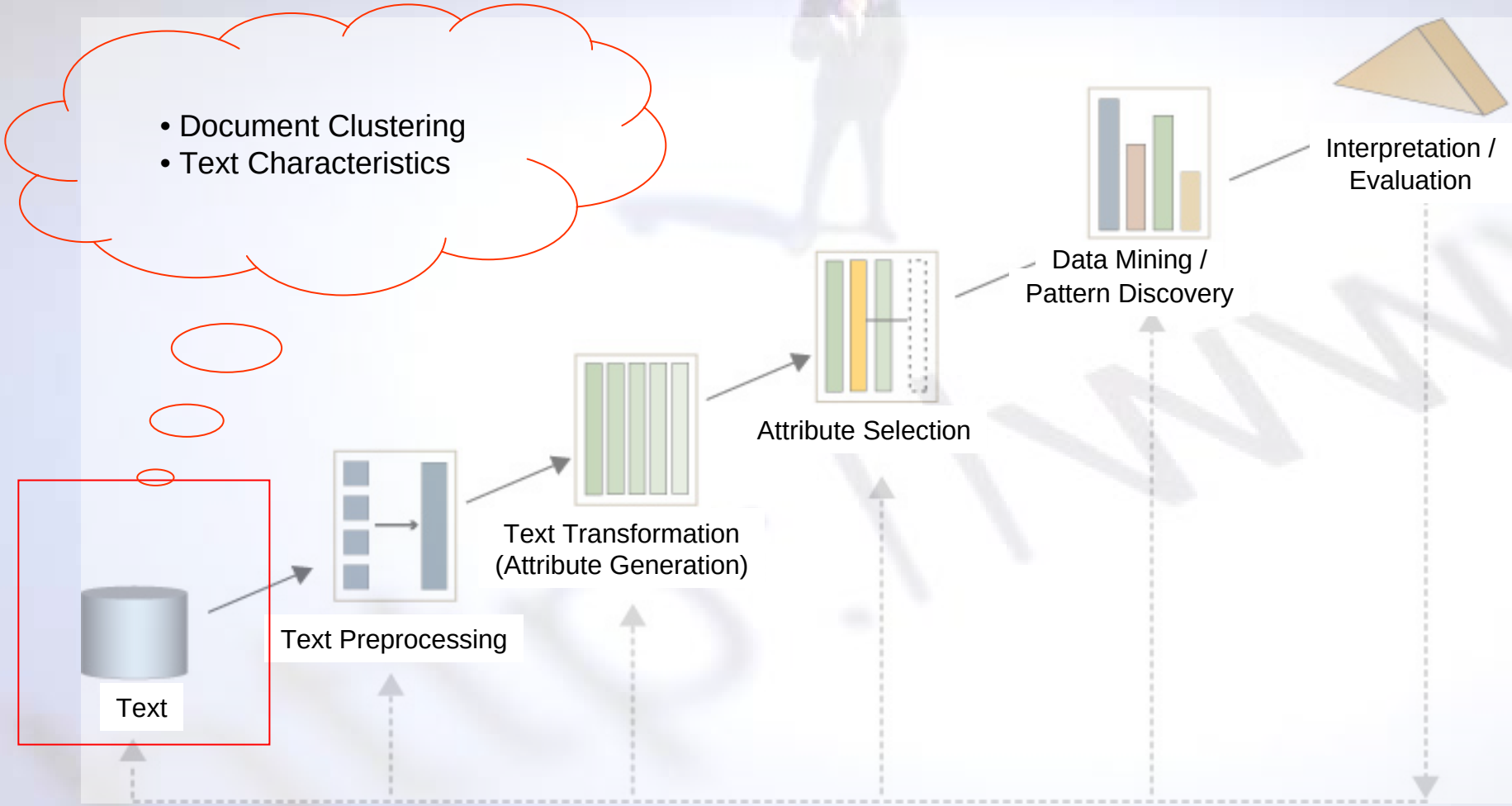
A classification of data mining and text data mining applications

	Finding Patterns	Finding Nuggets	
		Novel	Non-Novel
Non-textual Data	Standard Data Mining	?	Database Queries
Textual Data	Computational Linguistics	Real Text Data Mining	Information Retrieval

Hearst, M. *Untangling Text Data Mining.* In the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.


Data Mining
Information Retrieval
Statistics
Text Mining
Web Mining
Computational Linguistics &
Natural Language Processing

Text Mining Process



Document Clustering

- **Large volume of textual data**
  - **Billions of documents must be handled in an efficient manner.**
- **No clear picture of what documents suit the application.**
- **Solution: use Document Clustering (Unsupervised Learning).**
- **Most popular Document Clustering methods are:**
  - **K-Means clustering.**
  - **Agglomerative hierarchical clustering.**

Example: K-Means Clustering

- Given:
  - Set of documents (e.g. vector representation).
  - Suitable distance measure (e.g. cosine).
  - K (number of groups).
- For each of K groups initialize its centroid with a random document.
- While not converging
  - Each document is assigned to the nearest group (represented by its centroid).
  - For each group calculate new centroid (group mass point, average document in the group).

Text Characteristics

- **Several input modes**
  - **Text is intended for different consumers, i.e. different languages (human consumers) and different formats (automated consumers).**
- **Dependency**
  - **Words and phrases create context for each other.**

Text Characteristic contd…

- **Ambiguity**
  - **Word ambiguity.**
  - **Sentence ambiguity.**
- **Noisy data**
  - **Erroneous data.**
  - **Misleading (intentionally) data.**
- **Unstructured text**
  - **Chat room, normal speech, …**

Text Characteristic contd…

- High dimensionality (sparse input)
  - Tens of thousands of words (attributes).
  - Only a very small percentage is used in a typical document.
  - For example:
    - Top 2 words » 10-15% all word occurrences.
    - Top 6 words » 20% of all word occurrences.
    - Top 50 words » 50% of all occurrences.

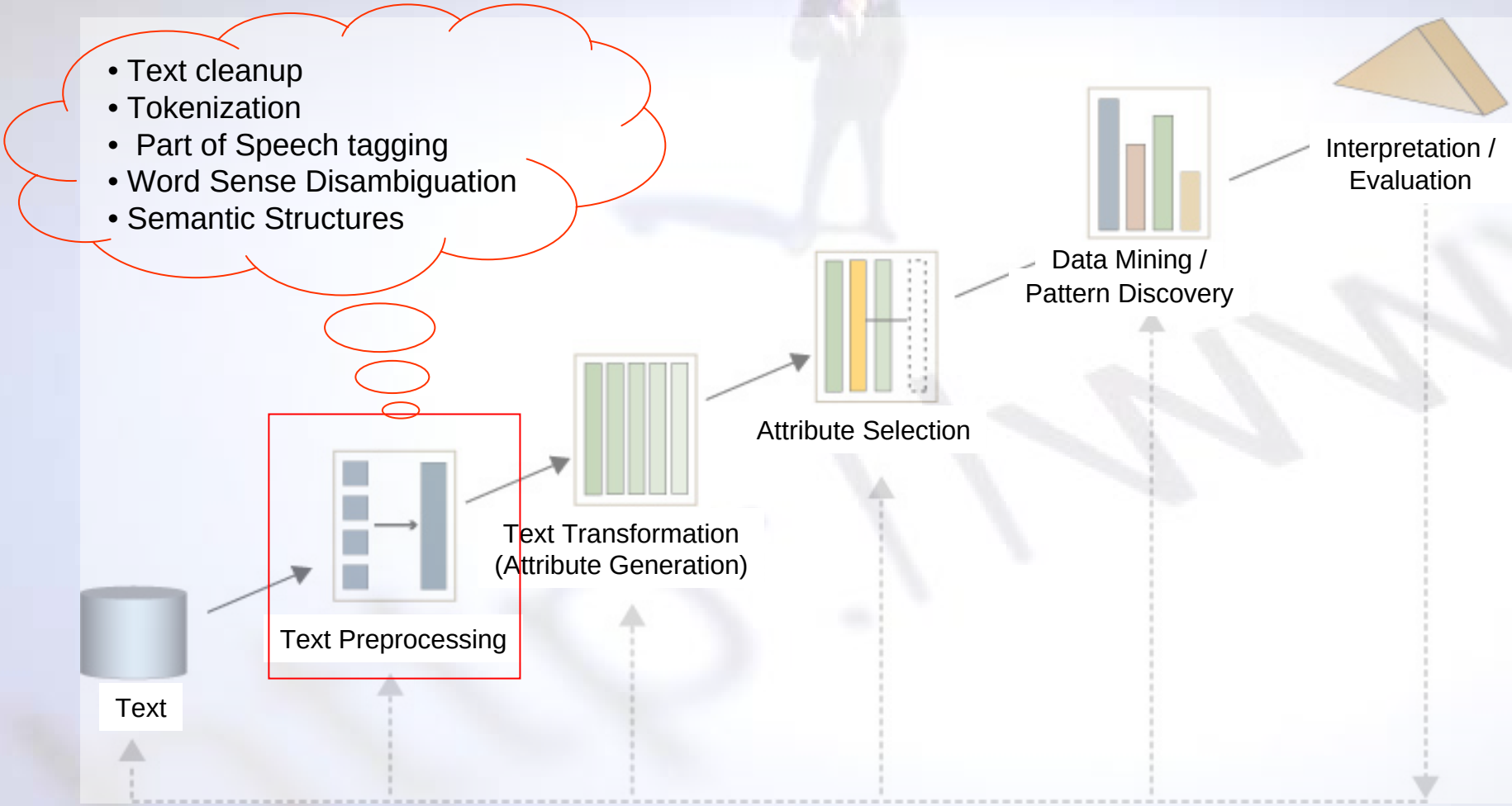
Word	Frequency
the	1,130,021
of	547,311
to	516,635
a	464,736
in	390,819
and	387,703
that	204,351
for	199,340

Word	Frequency
is	152,483
said	148,302
it	134,323
on	121,173
by	118,863
as	109,135
at	101,779
mr	101,679

Word	Frequency
with	101,210
from	96,900
he	94,585
million	93,515
year	90,104
its	86,774
be	85,588
was	83,398

WSJ87 collection (46,449 articles, 19 million term occurrences, 132 MB)

Text Mining Process



Text Preprocessing

- **Text cleanup**
  - **e.g., remove ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas, …**
- **Tokenization**
  - **Splitting up a string of characters into a set of tokens.**
  - **Need to deal with issues like:**
    - **Apostrophes, e.g., "John's sick", is it 1 or 2 tokens?**
    - **Hyphens, e.g., database vs. data-base vs. data base.**
    - **How should we deal with "C++", "A/C", ":-)", "…"?**
    - **Is the amount of white spaces significant?**

Text Processing contd…

- **Parts Of Speech tagging**
  - **The process of marking up the words in a text with their corresponding parts of speech.**
  - **Rule based**
    - **Depends on grammatical rules.**
  - **Statistically based**
    - **Relies on different word order probabilities.**
    - **Needs a manually tagged corpus for machine learning.**
- **Word Sense Disambiguation**
  - **Determining in which sense a word having a number of distinct senses is used in a given sentence.**
  - **“The king saw the rabbit with his glasses”**
    - **How many meanings?**

Text Processing again

- **Semantic Structures:**
  - **Two methods:**
    - **Full parsing: Produces a parse tree for a sentence.**
    - **Chunking with partial parsing: Produces syntactic constructs like Noun Phrases and Verb Groups for a sentence.**
  - **Which is better?**
    - **Producing a full parse tree often fails due to grammatical inaccuracies, novel words, bad tokenization, wrong sentence splits, errors in POS tagging, …**
    - **Hence, chunking and partial parsing is more commonly used.**

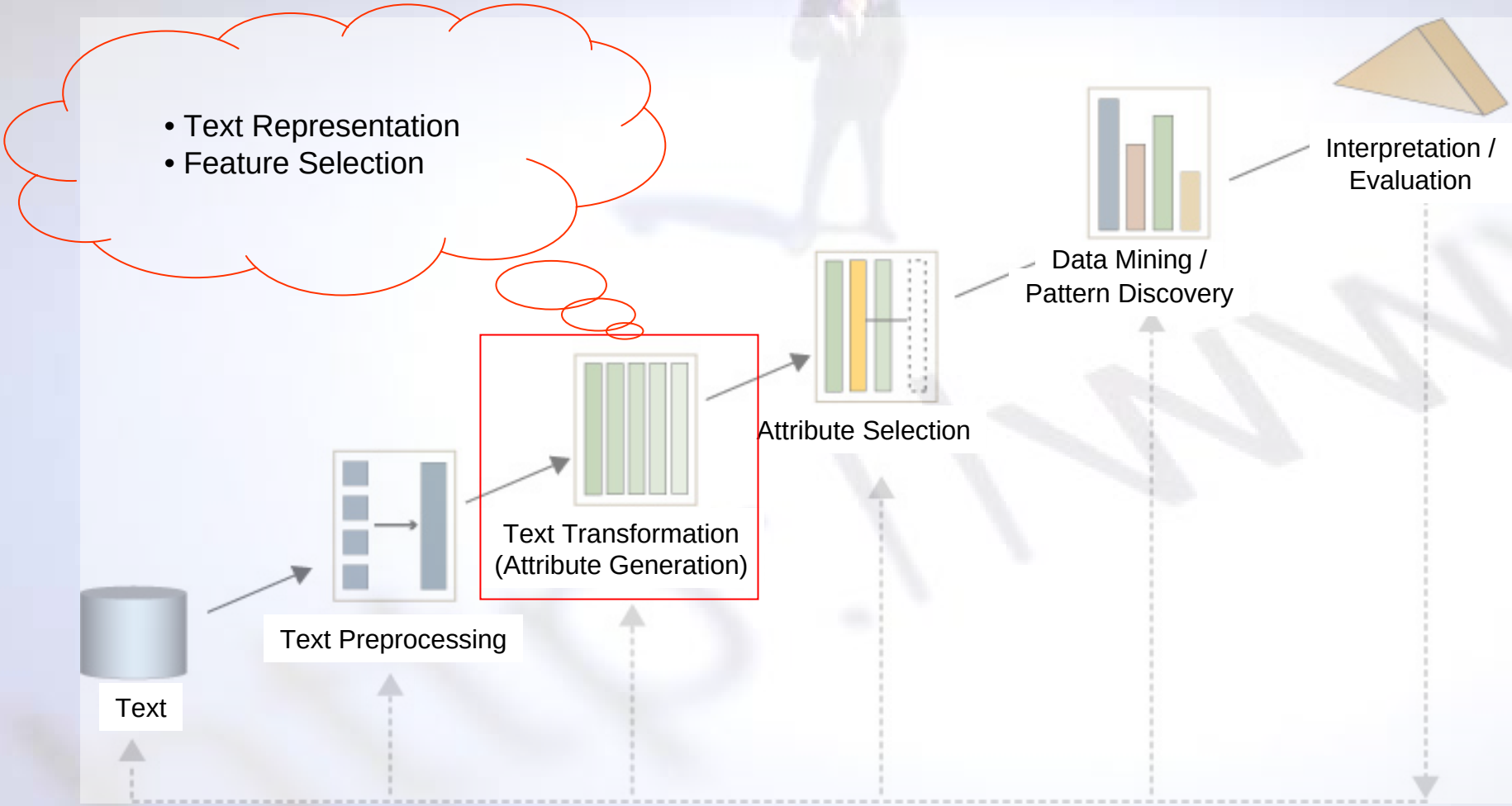
"I couldn't believe what I saw," said McNeill, who also discovered bomb-making instructions and detailed maps of U.S. landmarks in the cave. "On top of all the destruction these people had already unleashed, plans were underway to harass the American people with a merciless assault of offers for everything from discounts on home DSL lines to pre-approved, low-interest credit cards."

For all the evidence collected by the CIA, the "smoking gun" in the investigation may turn out to be an alleged Osama bin Laden motivational videotape, currently in the possession of CNN. The controversial tape, which has never aired on the cable network, is rumored to feature bin Laden urging his followers to think positive and believe in the quality of the product they are pitching, closing on the grim slogan "Smile And Dial."

ype	Set	Start	End	Features
P	Default	3582	3596	{DET="", MOD="", HEAD="Guantanamo Bay "}
P	Default	776	791	{DET="the ", MOD="dinner ", HEAD="hour "}
P	Default	2259	2262	{DET="", MOD="", HEAD="out "}
P	Default	1806	1807	{DET="", MOD="", HEAD="I "}
P	Default	3849	3852	{DET="", MOD="", HEAD="one "}
P	Default	987	996	{DET="The ", MOD="", HEAD="video "}
P	Default	1487	1494	{DET="", MOD="", HEAD="McNeill "}
P	Default	2280	2318	{DET="", MOD="Osama bin Laden motivational ", HEAD="videotape "}
P	Default	894	910	{DET="", MOD="money ", HEAD="laundering "}

Witte, R. *Prelude Overview: Introduction to Text Mining Tutorial*. EDBT, 2006.

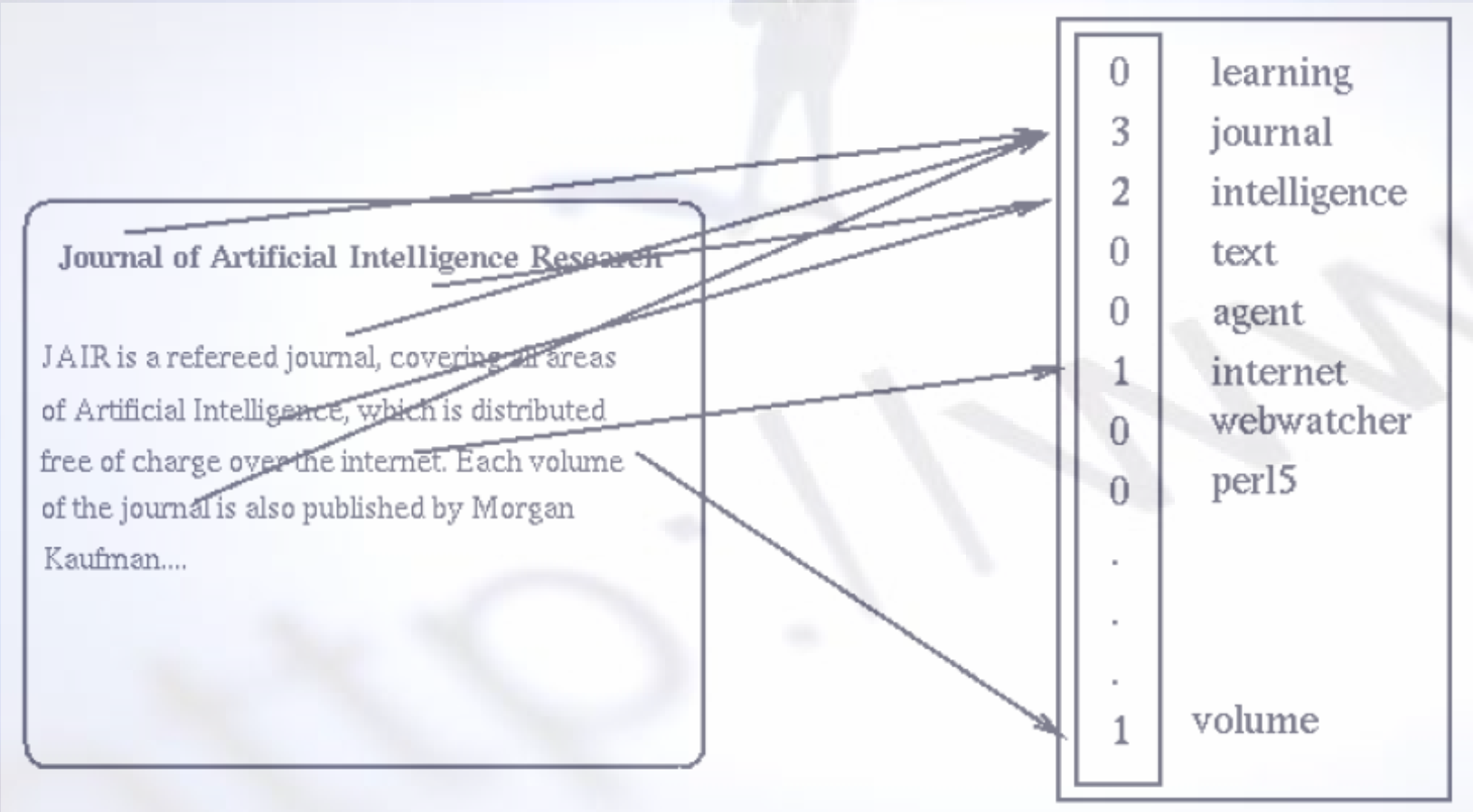
Text Mining Process



Attribute Generation

- **Text Representation:**
  - **Text document is represented by the words (features) it contains and their occurrences.**
  - **Two main approaches of document representation**
    - **“Bag of words”.**
    - **Vector Space.**
- **Feature Selection:**
  - **Which features best characterize a document?**
- **Actual Attribute Generation:**
  - **We use a classifier to automatically generate labels (attributes) from the features we feed into it.**

“Bag of words” Document Representation



Grobelnik, M. and Mladenic, D. Text-Mining Tutorial. In the Proceeding of Learning Methods for Text Understanding and Mining, Grenoble, France, January 26 – 29, 2004.

Word Weighting

- **In "Bag of words" representation each word is represented as a separate variable having numeric weight.**
- **The most popular weighting schema is normalized word frequency *tfidf*:**

$$tfidf\ (w) = tf\ \cdot \log(\frac{N}{df\ (w)})$$

- ***tf(w)* –term frequency (number of word occurrences in a document)**
- ***df(w)* –document frequency (number of documents containing the word)**
- ***N* –number of all documents**
- ***tfidf(w)* –relative importance of the word in the document**

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Vector Space Document Representation

- **TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.**
- **[RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171] [ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119] [DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102] [DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080] [MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070] [REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064] [OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056] [SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041] [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]**

Feature Selection

- What is feature selection?
  - **Select just a subset of the features to represent a document.**
  - **Can be viewed as creating an improved text representation.**
- Why do it?
  - **Many features have little information content**
    - **e.g. stop words.**
  - **Some features are misleading**
  - **Some features are redundant**
    - **Independence assumptions result in double-counting**
  - **Some algorithms work better with small feature sets**
    - **e.g. because they create complex classifiers…**
      …so the space of possible classifiers is very large

Feature Selection contd…

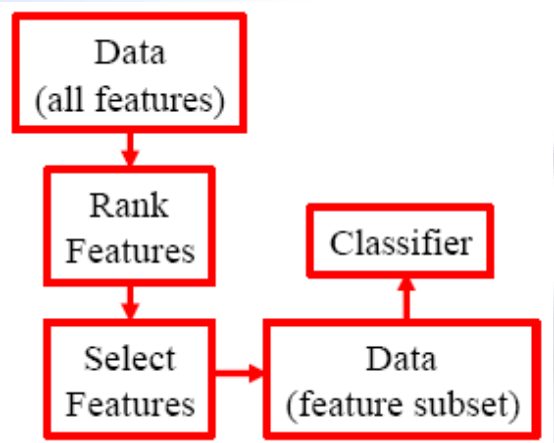
- **Stop words removal**
  - **The most common words are unlikely to help text mining, e.g., “the”, “a”, “an”, “you” …**
- **Stemming**
  - **Identifies a word by its root.**
  - **Reduce dimensionality (number of features).**
  - **e.g. flying, flew → fly**
  - **Two common algorithms :**
    - **Porter’s Algorithm.**
    - **KSTEM Algorithm.**

Feature Selection contd…

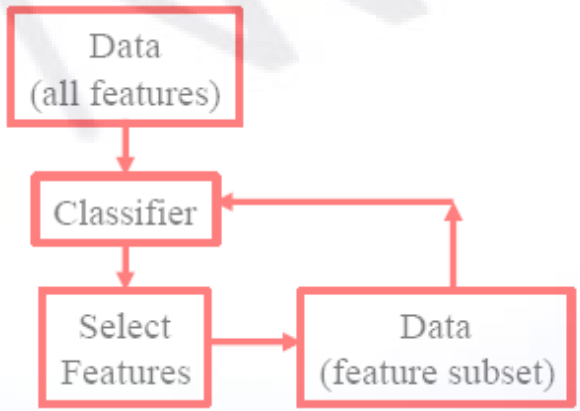
- **Stemming Examples**
  - **Original Text**
    - **Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals.**
  - **Porter Stemmer (stop words removed)**
    - **market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem**
  - **KSTEM (stop words removed)**
    - **marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic**

Two Approaches to Feature Selection

- Select features before using them in a classifier
  - **Requires a feature ranking method.**
  - **Many choices.**



- Select features based on how well they work in a classifier
  - **The classifier is part of the feature selection method.**
  - **Often an iterative process.**



Callan, J. A Course on Text Data Mining. Carnegie Mellon University, 2004.

Two Approaches to Feature Selection contd…

Select Before Use

- Evaluation of features is independent of classifier
  - **Many choices.**
- Evaluate each feature once.
- Lower computational costs
  - **Simpler algorithms.**
- Less effective at identifying redundant features
  - **Features are usually evaluated individually.**
  - **Redundancy can be a classifier-specific property.**

Select Based On Use

- Evaluation of features by how they perform in actual use
  - **A more tailored approach.**
- Evaluate features iteratively.
- Higher computational costs
  - **Must train the classifier.**
- Can be more effective
  - **But effectiveness depends on classifier's ability to evaluate features.**

Actual Attribute Generation

- Attributes generated are merely labels of the classes automatically produced by a classifier on the features that passed the feature selection process.
- The next step is to populate the database that results from above.
- The figure on the next slide depicts this process.

Attribute Generation

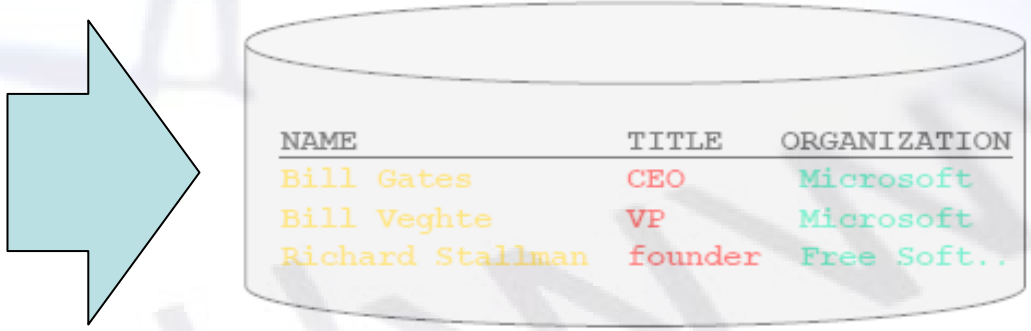
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

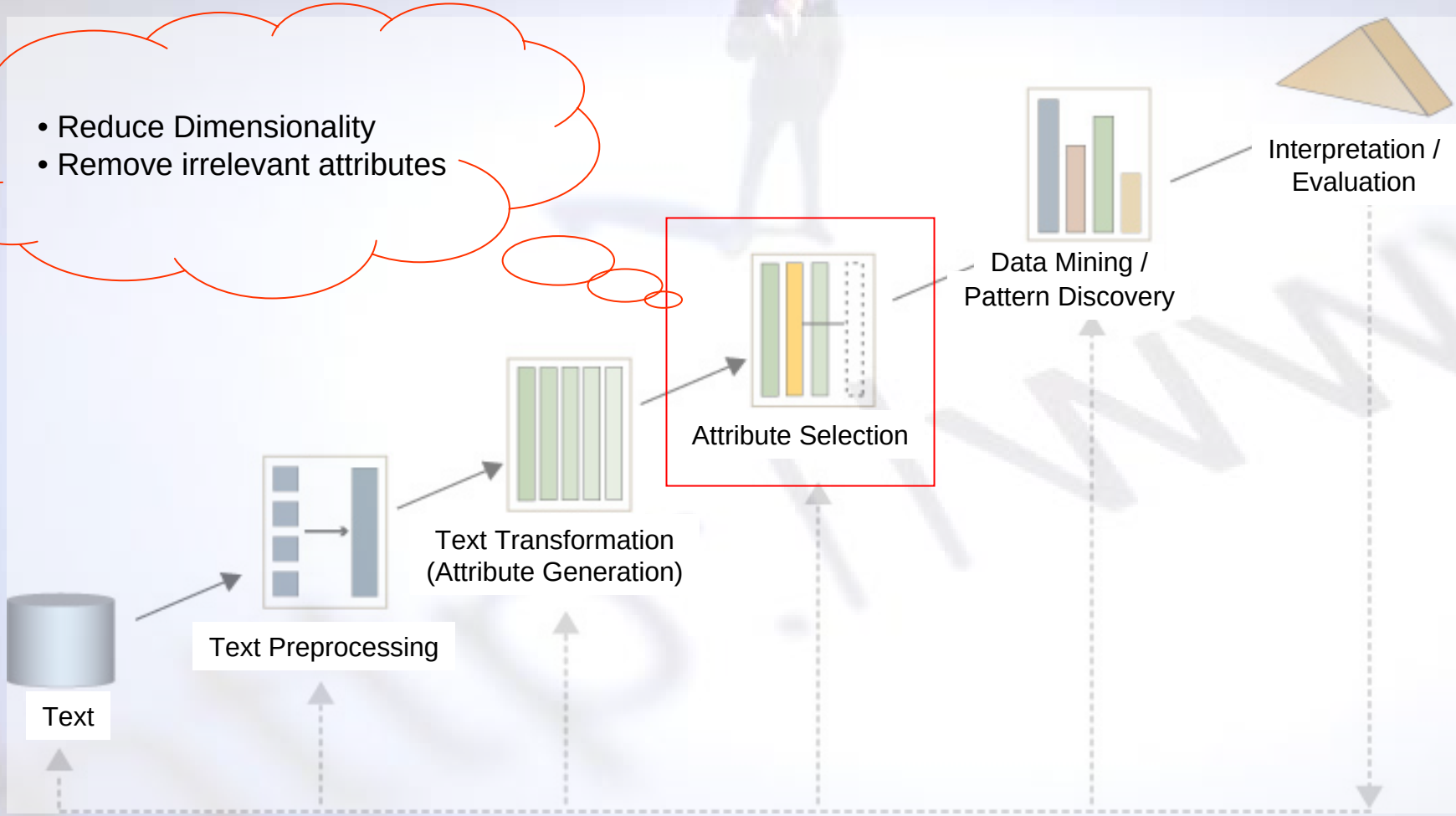
Richard Stallman, founder of the Free Software Foundation, countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Grobelnik, M. and Mladenic, D. Text-Mining Tutorial. In the Proceeding of Learning Methods for Text Understanding and Mining, Grenoble, France, January 26 – 29, 2004.

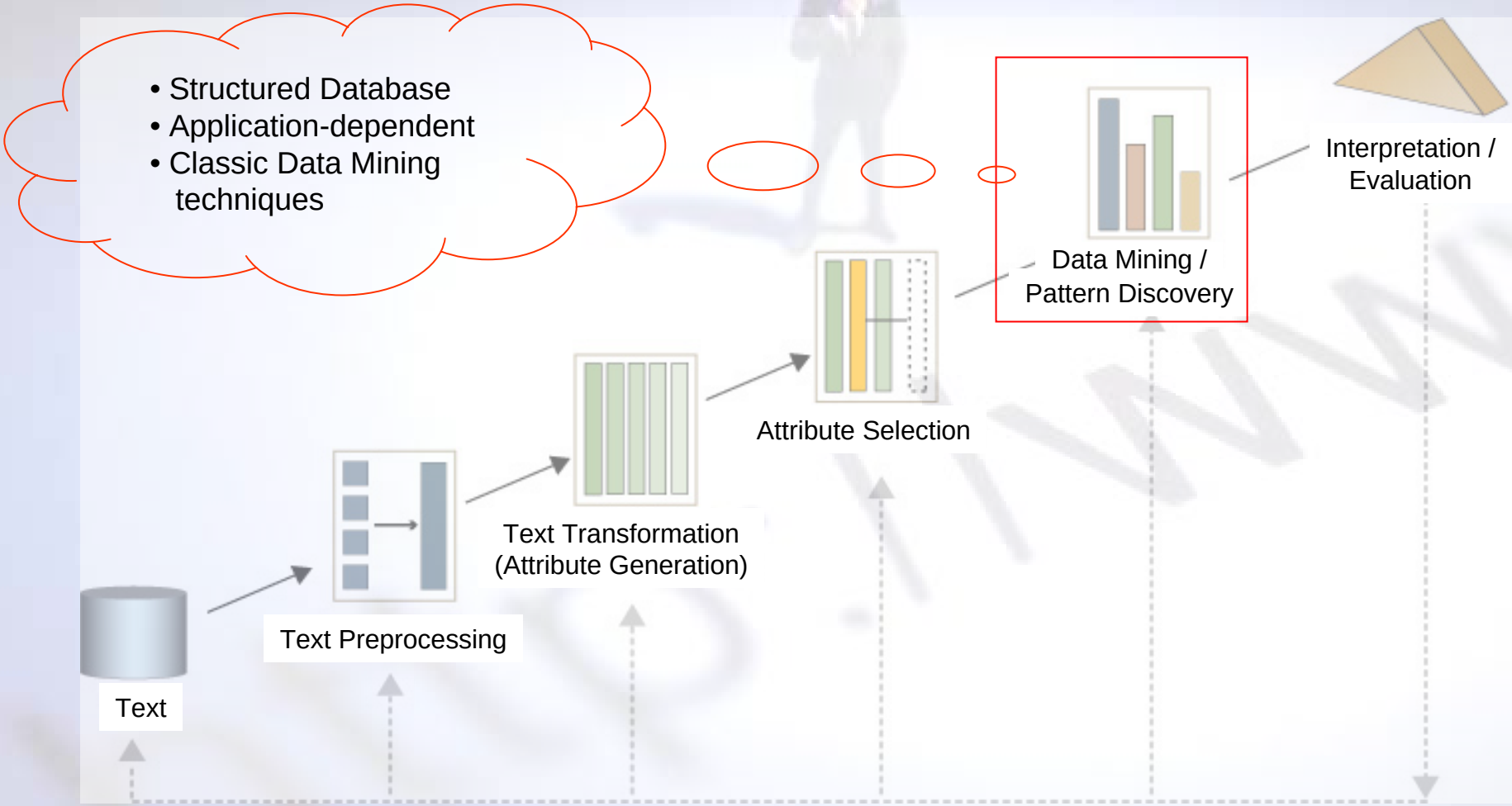
Text Mining Process



Attribute Selection

- Further reduction of dimensionality
  - Learners have difficulty addressing tasks with high dimensionality.
  - Scarcity of resources and feasibility issues also call for a further cutback of attributes.
- Irrelevant features
  - Not all features help!
    - e.g., the existence of a noun in a news article is unlikely to help classify it as “politics” or “sport”.

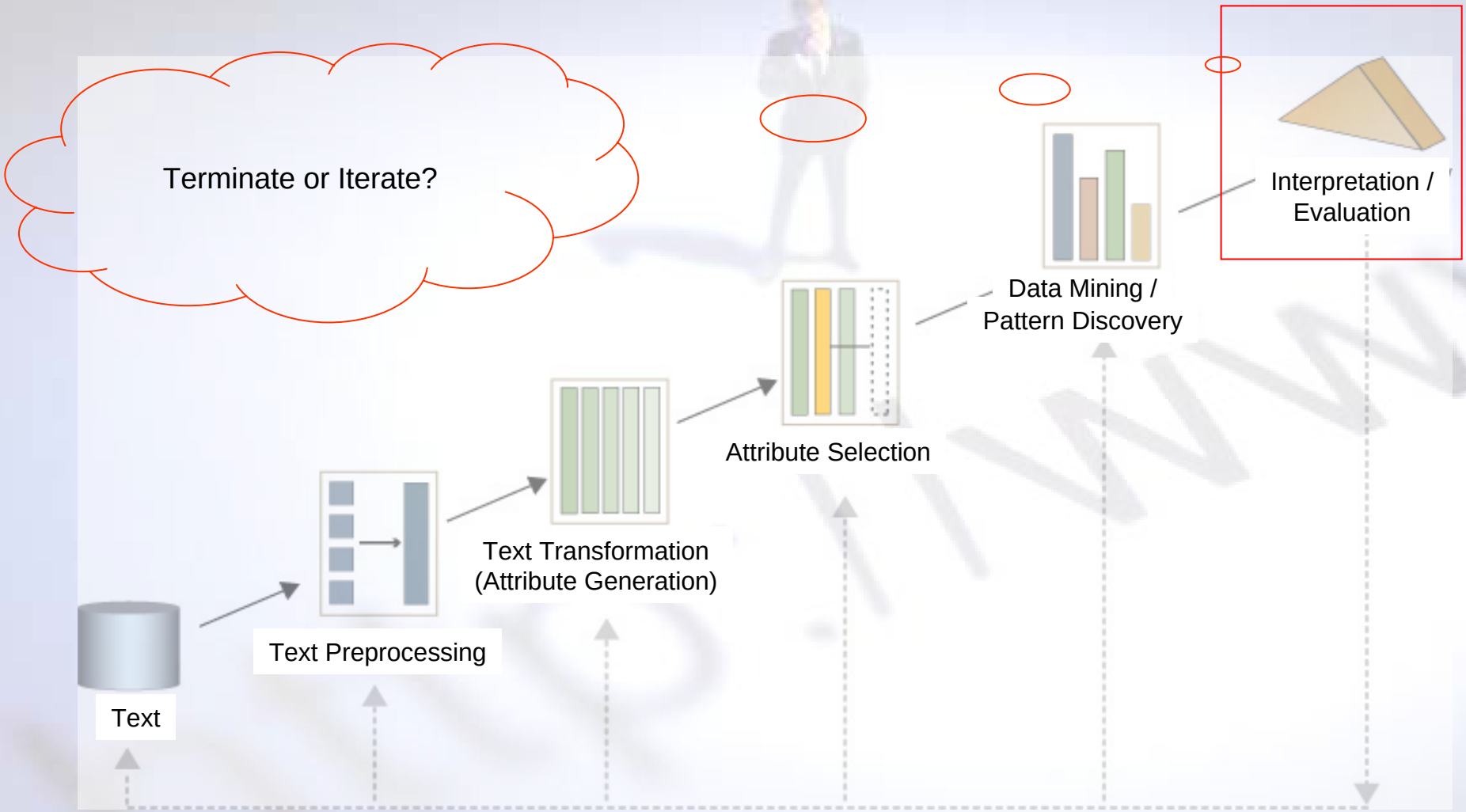
Text Mining Process



Data Mining

- At this point the Text mining process merges with the traditional Data Mining process.
- Classic Data Mining techniques are used on the structured database that resulted from the previous stages.
- This is a purely application-dependent stage.

Text Mining Process



Interpretation & Evaluation

- **What to do next?**
  - **Terminate**
    - **Results well-suited for application at hand.**
  - **Iterate**
    - **Results not satisfactory but significant.**
    - **The results generated are used as part of the input for one or more earlier stages.**

Using text in Medical Hypothesis Discovery

- For example, when investigating causes of migraine headaches, Don Swanson extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:
  - Stress is associated with migraines.
  - Stress can lead to loss of magnesium.
  - Calcium channel blockers prevent some migraines.
  - magnesium is a natural calcium channel blocker.
  - Spreading Cortical Depression (SCD) is implicated in some migraines.
  - High leveles of magnesium inhibit SCD.
  - Migraine patients have high platelet aggregability.
  - Magnesium can suppress platelet aggregability.

Using text in Medical Hypothesis Discovery

- These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links.
- The hypothesis has to be tested via non-textual means, but the important point is that a new, potentially plausible medical hypothesis was derived from a combination of text fragments and the explorer's medical expertise.
- According to [Swanson1991], subsequent study found support for the magnesium-migraine hypothesis [Ramadan1989].

Linguistic Profiling for Author Recognition and Verification

Hans van Halteren
Univ. of Nijmegen, The Netherlands

42nd Annual Meeting of the
Association for Computational Linguistics
Forum Convention Centre Barcelona. July 21-26, 2004.

References

- [1] H. Baayen, H. V. Halteren, A. Neijt, and F. Tweedie. An experiment in authorship attribution. 6th JADT, 2002.
- [2] J. Burrows. Word patterns and story shapes: the statistical analysis of narrative style. Literary and linguistic Computing, 2:61-70, 1987.
- [3] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. Applied Intelligence, 19(1-2):109-123, 2003.
- [4] P. Juola and H. Baayen. A controlled-corpus experiment in authorship identication by cross-entropy. Literary and Linguistic Computing, 2003.
- [5] D. I. Holmes, M. Robertson, and R. paez. Stephen crane and the new-york tribune: A case study in traditional and non-traditional authorship attribution. Computers and the Humanities, 35(3):315-331, 2001.

References

- [6] H. Baayen, H. V. Halteren, and F. Tweedie. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3):121-132, 1996.
- [7] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, pages 158-164, 1999.
- [8] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2):193-214, 2001.
- [9] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In Pacific Association for Computational Linguistics, pages 256-264, 2003.
- [10] F. Peng, D. Schuurmans, V. Keselj, and S. Wang. Language independent authorship attribution using character level language models. In 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL, 2003.

References

- [11] Harald Baayen, Hans van Halteren, Anneke Neijt, and Fion Tweedie. 2002. An Experiment in Authorship Attribution. Proc. JADT 2002, pp. 69-75.
- [12] Ton Broeders. 2001. Forensic Speech and Audio Analysis, Forensic Linguistics 1998-2001 – A Review. Proc. 13th Interpol Forensic Science Symposium, Lyon, France.
- [13] C. Chaski. 2001. Empirical Evaluations of Language-Based Author Identification Techniques. *Forensic Linguistics* 8(1): 1-65.
- [14] Peter Arno Coppen. 2003. Rejuvenating the Amazon parser Poster presentation CLIN2003, Antwerp, Dec. 19, 2003.
- [15] David Holmes. 1998. Authorship attribution. *Literary and LinguisticComputing* **13(3):**111-117.
- [16] P. C. Uit den Boogaart. 1975. *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema & Holkema, Utrecht.

References

- [17] F. Mosteller, and D.L. Wallace. 1984. *Applied Bayesian and Classical Inference in the Case of the Federalist Papers* (2nd edition). Springer Verlag, New York.
- [18] Hans van Halteren, Jakub Zavrel, and Walter Daelemans. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics* 27(2):199-230.
- [19] Hans van Halteren and Nelleke Oostdijk, 2004. Linguistic Profiling of Texts for the Purpose of Language Verification. Proc. COLING 2004.
- [20] Hans van Halteren, Marco Haverkort, Harald Baayen, Anneke Neijt, and Fiona Tweedie. To appear. New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*.

Abstract

- Several approaches are available for authorship verification and recognition
- We introduce a new technique – Linguistic Profiling
- We achieved 8.1% false accept rate (FAR) with false reject rate (FRR) 0% for verification
- Also 99.4% 2-way recognition accuracy

Introduction

- Authorship attribution is the task of deciding who wrote a document
- A set of documents with known authorship is used for training
- The problem is to identify which of these authors wrote unattributed documents
- Typical uses include-
  - Plagiarism detection
  - Verify claimed authorship

Introduction

- An interesting site dedicated to the proposition that “Shakespeare Wrote Shakespeare” (www.shakespeareauthorship.com)
  - Identify emails, newsgroup messages or a piece of intelligence
  - And more….
- A variety of approaches has been proposed

Introduction

- Lexical methods [1, 2, 3, 4, 5]
- Syntactic or grammatical methods [6, 7, 8]
- Language model methods [9, 10]
- These approaches vary in evidence or features extracted from documents and in classification methods applied (Bayesian network, Nearest-neighbor methods, Decision trees, etc.)

Introduction

- **Problems are divided into several categories:**
  - **Binary Classification: each of the documents is known to have been written by one of two authors**
  - **Multi-class Classification: documents by more that two authors are provided**
  - **One-class Classification: some documents are by a single author, others unspecified**

Features Used

- Usually words in the document
- But the task is different from document classification
- Authors writing on same topics may share many common words. So it may be misleading
- So, we need style markers rather than content markers

Features Used

- **If words are used, *function words* are more interesting**
- **These are words such as prepositions, conjunctions or articles**
- **They have little semantic content but are markers of writing style**
- **Less common function words are more interesting, e.g. “whilst” or “notwithstanding” are rarely used, therefore a good indicator of authorship**

Features Used

- Other aspects of text such as word length or sentence length can also be used as features
- Richer features are available through NLP or more complicated statistical modeling
- They are mainly syntactic annotation (like finding noun phrases)

A Big Challenge

- No benchmarking dataset available to make a fair comparison among the methods proposed
- Everyone claims to be winner

Quality Measures

- **Basic Measures:**
  - **False Accept Rate (FAR)**
  - **False Reject Rate (FRR)**
- **When FAR goes down, FRR goes up**
- **The behavior of the system can be shown by one of several types of FAR/FRR curve-**
  - **FAR vs FRR plot (Receiver Operating Characteristic curve)**

Quality Measures

- Equal Error Rate (EER), i.e. FAR = FRR
- FAR when FRR = 0 (no false accusations)
- FRR when FAR = 0 (no guilty unpunished)

- We would like to measure the quality of the system with the FAR at the threshold at which the FRR becomes zero
- Because in situations like plagiarism detection, we don't want to accuse someone unless we are sure

Test Corpus (Collection of Text)

- 8 students
- 9 text from each student
- Fixed subjects (3 argumentative, 3 descriptive, 3 fiction)
- About 1,000 words per text

Profiling

- A profile vector is constructed for each author from a large number of linguistic features
- The vector contains the standard deviations of the counts of features observed in the profile reference corpus
- This vector will be used like a fingerprint of the author

Authorship Score Calculation

- The system has to decide if an unattributed text is written by a specific author, on the basis of the attributed texts
- System's ability to make this distinction was tested by means of a 9-fold cross validation experiment
- During a run, the system only knows whether a text is written by a specific author or not by this author

Authorship Score Calculation

- Author profile = mean of the profiles for the known texts
- Text verification score = distance measure (text profile to author profile)
- Distance measure =

$$\Delta_T = (\sum | T_i - A_i |^D | T_i |^S)^{1/(D+S)}$$

- Ti = value for the ith feature for the text sample
- Ai = value for the ith feature for the author
- D, S = weighting factors

Authorship Score Calculation

- **This measure is then transformed into a score by the formula**

$$Score_T = (\sum | T_i |^{(D+S)})^{1/(D+S)} - \Delta_T$$

- **The higher the score, the more the similarity between text sample profile and author profile**

Results with Lexical Features

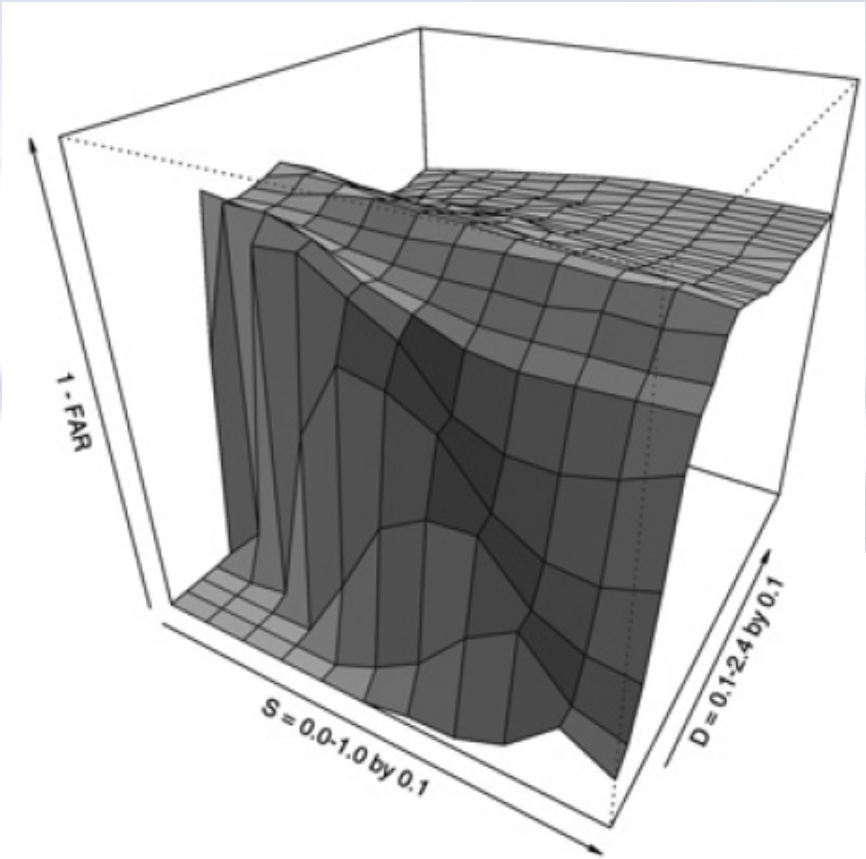
- **FAR when FRR=0**

as function of D

and S

- **Best result (15%)if**

D=0.60 and S=0.15



Results with Syntactic Features

- Amazon Parser (http://lands.let.kun.nl/~dreumel/amacas.en.html) is used to extract syntactic features (details about the parser is in Dutch)
- The size of the feature vector is about 900k counts
- Best result is 25% at D = 1.3, S = 1.4
- Worse than lexical feature analysis

…So Combine the Features

- For now, combination means addition
- We add the two scores from two analysis
- The combination of the best two individual systems leads to an FAR of 10.3% (with FRR = 0)
- But the best combination produces 8.1%

Comparison with Other Methods

	2-way errors /504	2-way percent correct	8-way errors /72	8-way percent correct
50 function words, PCA		± 50%		
followed by LDA		± 60%		
LDA with cross-sample entropy weighting		± 80%		
all tokens, WPDV modeling		97.8%		
Lexical	6	98.8%	5	93%
Syntactic	14	98.2%	10	86%
Combined	3	99.4%	2	97%
Lexical (renorm.)	1	99.8%	1	99%
Syntactic (renorm.)	4	99.2%	3	96%
Combined (renorm.)	0	100.0%	0	100%

Hans van Halteren, Linguistic Profiling for Author Recognition and Verification. 42nd Annual Meeting of the Association for Computational Linguistics Forum Convention Centre Barcelona. July 21-26, 2004.

Concluding Remarks

- The first issue that can be addressed is “parameter setting”
- There is no dynamic parameter setting scheme
- Results with other corpora might also provide interesting results
- Different kinds of feature selection may provide better results……

ARROWSMITH

discovery from complementary literatures

Don R. Swanson(d-swanson@uchicago.edu)
The University of Chicago.

Project Links-
http://kiwi.uchicago.edu/
http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

References-
Swanson DR, Smalheiser NR, Torvik VI. Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). JASIST 2006, in press.

Overview

- Extends the power of a MEDLINE search.
- Information developed in one area of research can be of value in another without anyone being aware of the fact.
- Direct vs indirect connections between two literatures.
- ABC model- key B-terms (words and phrases) in titles that are common to two disjoint sets of articles, A and C.

Overview

- **ARROWSMITH begins with a question concerning the connection between two entities for which the relation is to be determined.**
- **Conventional searching provides no answer .**
- **A→X and X→C cannot be discovered by a conventional database search techniques without prior knowledge of X.**

Stages– Preparatory Steps

- Search MEDLINE for the intersection "A AND C" for any direct relation.
- For an indirect relation, proceed.
- Search MEDLINE title-word search for the word or term denoted by C and by A separately and the save the files with the summary format.
- Title-word searching may be enhanced by including subject-headings as well.

Stage 1

- Upload both the files to ARROWSMITH.
- A list (called the "B-LIST") of MeSH terms common to both of files is produced
- If the input format includes Medical Subject Headings, these also participate in the matching process.
- Title terms ranked according to the number of MeSH terms that are shared by the A and C titles.
- Each of the title terms is a potential candidate for the mysterious "X" mentioned above.
- Title-based list in general should be edited by the user.

Stage 2

- Delete entries from the B-LIST produced by STAGE 1.
- Initial B-list includes terms that are not useful.
- Medical Subject Headings (MeSH) have been integrated into the matching process and the title display for ranking the B-list terms.
- All terms having rank 0 are automatically eliminated from B-lists, thus reducing the need for manual editing.
- B-list can be edited forming groups; it is helpful to bring together synonyms and related terms

Stage 3

- Permits repeated browsing of results formed in all other stages -- B-list, title files and the ranked A-list.
- Each B-LIST is a series of links.
- Clicking on any B-term "X" results in displaying the corresponding titles that contain both A and "X" and, next to these, titles that contain "X" and C.
- Iterative process –user can go back to stage 2

Stages 4 and 5

- From the broad-category titles, Stage 4 constructs a list of individual terms, within those titles, and ranks them according to the number of different bridging terms, B.
- The A-list can be edited either by deleting terms, or by grouping terms. If the resulting A-list seems unmanageably large, go back to Stage 2 and delete unwanted terms from the originalB-list.
- The last stage permits you to continue to edit the A-list produced in Stage 4. (If you wish to start the editing over from the beginning, then repeat Stage 4; if you wish only to inspect or browse results, go to Stage 3

Author_ity

- Provides a pairwise ranking of articles by similarity to a given index paper, across 9 different attributes and based on that calculate the Prm value.
- PrM value -- estimate of the probability that the paper is authored or co-authored by the same individual as the index paper.
- PrM > 0.5 will correspond to the same author, and the higher the value, the greater the chance that they share the same author.

Ranking Startegy Used

- Resulting number of key B-terms might be in the order of millions.
- Solution address on two two fronts
  - Trying to improve the search strategies used in creating files A and C.
  - Filtering and organizing the B-list
- Medical Subject Headings (MeSH) play a key role on both fronts.

Target

- Any B-term that is judged by the user to be of scientific interest because of its relationship to both the A and C literatures is called a "target".
- Target terms potentially may lead to literature-based discovery.
- ARROWSMITH provides a link from each B-term to the A and C titles from which it was extracted, and so helps the user assess whether it might qualify as a target.

Stop words

- Stop words -- lists of words to be excluded because they are predictably of no interest
- Compiled by selecting words from a composite, frequency ranked B-list automatically created.
- Medical Subject Headings used to index Medline records are also filtered using a MeSH stop words of 4900 terms.
- MeSH terms within top-level or second-level MeSH categories form the main 4000-term core of the stoplist.

Ranking Strategy

- Usefulness of B-term depends ultimately on the contents of the articles within which that term co-occurs with A and with C.
- B-list Ranking using MeSH terms.
- Identify, automatically, subsets of B-terms that are likely to have higher target density, and are given a higher rank, than other subsets.
- Interpreting that context and its usefulness in suggesting new relationships requires in general, expert knowledge and human judgment.

Ranking Strategy

- Each B-word corresponds to a small set of records from the A-file and from the C-file.
- MeSH terms in these records provide context make it easier for the viewer to assess an A-C relationship.
- greater density of MeSH terms that the corresponding AB and BC records have in common, more possibility of suggestive relationship between A and C.
- We will define a ranking formula based on Mesh terms now.

Weightage formula for ranking

For a given B list term

- {AB} = subset of records in A containing that title-term.
- {BC} = subset of records in C containing that title-term.
- nAB = number of records in {AB}
- nBC = number of records in {BC}
- ncom = the number of unique subject headings that {AB} and {BC} have in common.
- weight for a given title B-term =100*ncom/(nAB*nBC).

Example

- AB title is about magnesium and ischemia.
- BC title is on ischemia and migraine.
- Possibility of a magnesium-migraine connection via the B-term "ischemia" is likely to be greater if the two uses of "ischemia" are in the same context.
- Both cerebrovascular) rather than in different contexts (such as ABcardiovascular and BC-cerebrovascular).
- Corresponding MeSH terms displayed to the searcher, help to resolve this point.

THANK YOU