10 Text Mining

Munawar, PhD

Definition

- Text mining also is known as Text Data Mining (TDM) and Knowledge Discovery in Textual Database (KDT).[1]
- A process of identifying novel information from a collection of texts (also known as a corpus). [2]

Data Mining vs. Text Mining

Data Mining

- process directly
- Identify causal relationship
- Structured numeric transaction data residing in rational data warehouse

Text Mining

- Linguistic processing or natural language processing (NLP)
- Discover heretofore unknown information[2]
- Applications deal with much more diverse and eclectic collections of systems and formats[4]

Disputation

- Hearst: non-novel, novel.
  - Text mining is not a simple extension of data mining applied to unstructured database.
  - Text mining is the process of mining precious nuggets of ore from a mountain otherwise worthless rock.
- Kroeze: non-novel, semi-novel, and novel
  - Non-novel: data/information retrieval
  - Semi-novel: knowledge discovery (standard data-mining, metadata mining, and standard text mining)
  - Novel: intelligent text mining

Confusion

- Is text mining the same as information extraction? No!
- Information Extraction (IE)
  - Extract facts about pre-specified entities, events or relationships from unrestricted text sources.
  - No novelty: only information is already present is extracted.

Two Foundations

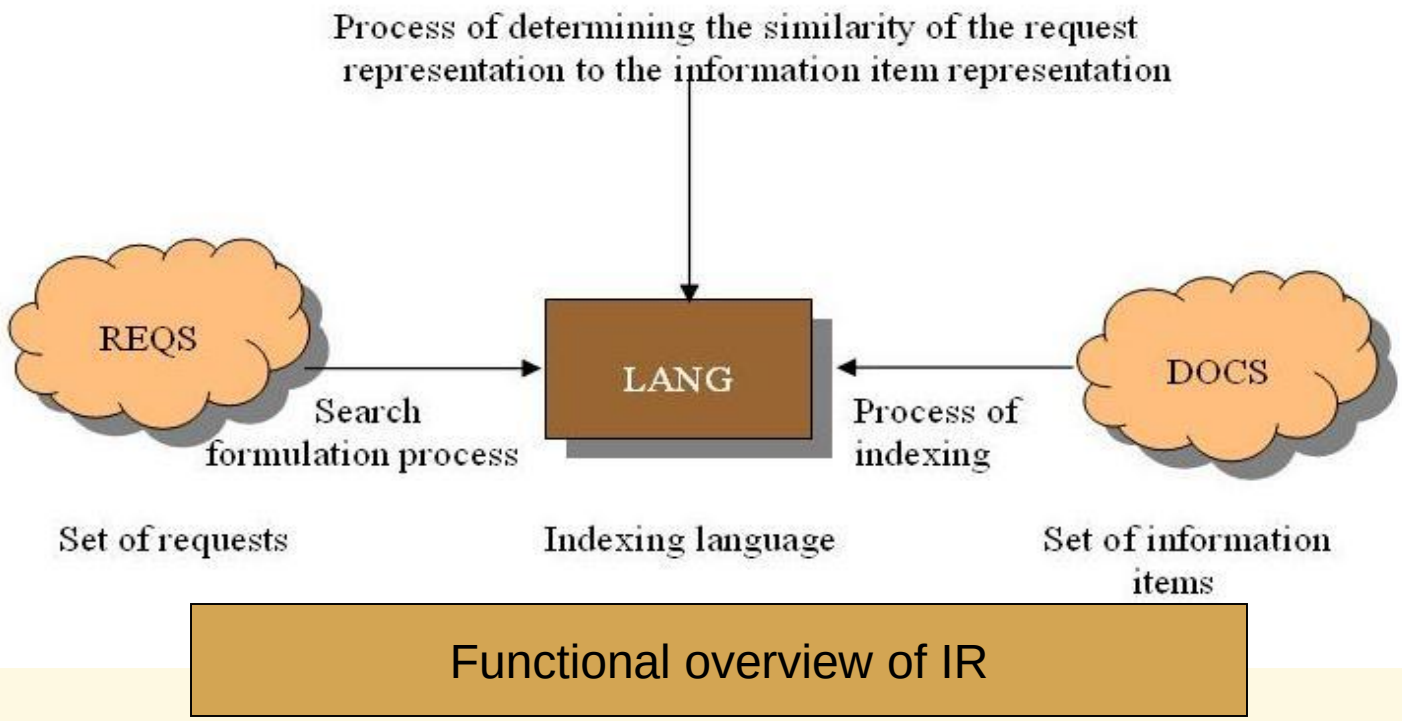
- Information Retrieval (IR)
- Artificial Intelligence (AI)

Information Retrieval

- The science of searching for
  - Information in documents
  - Documents themselves
  - Metadata which describe documents
  - Text, sound, images or data, within database: relational stand-alone database or hypertext networked databases such as the Internet or intranets.

Information Retrieval

- Gerard Salton



Functional overview of IR

Application I

- Semi-novel
- Text clustering: group similar documents for further examination -> create thematic overviews of text collections Issues:
  - Information needs is vague
  - Even if a topic were available, the words used to describe it may not be known to the user
  - The words used to describe a topic may not be those used to discuss the topic and may thus fail to appear in articles of interest.
  - Even if some words used in discussion of the topic were available, documents may fail to use precisely those words. [5]

Text Clustering

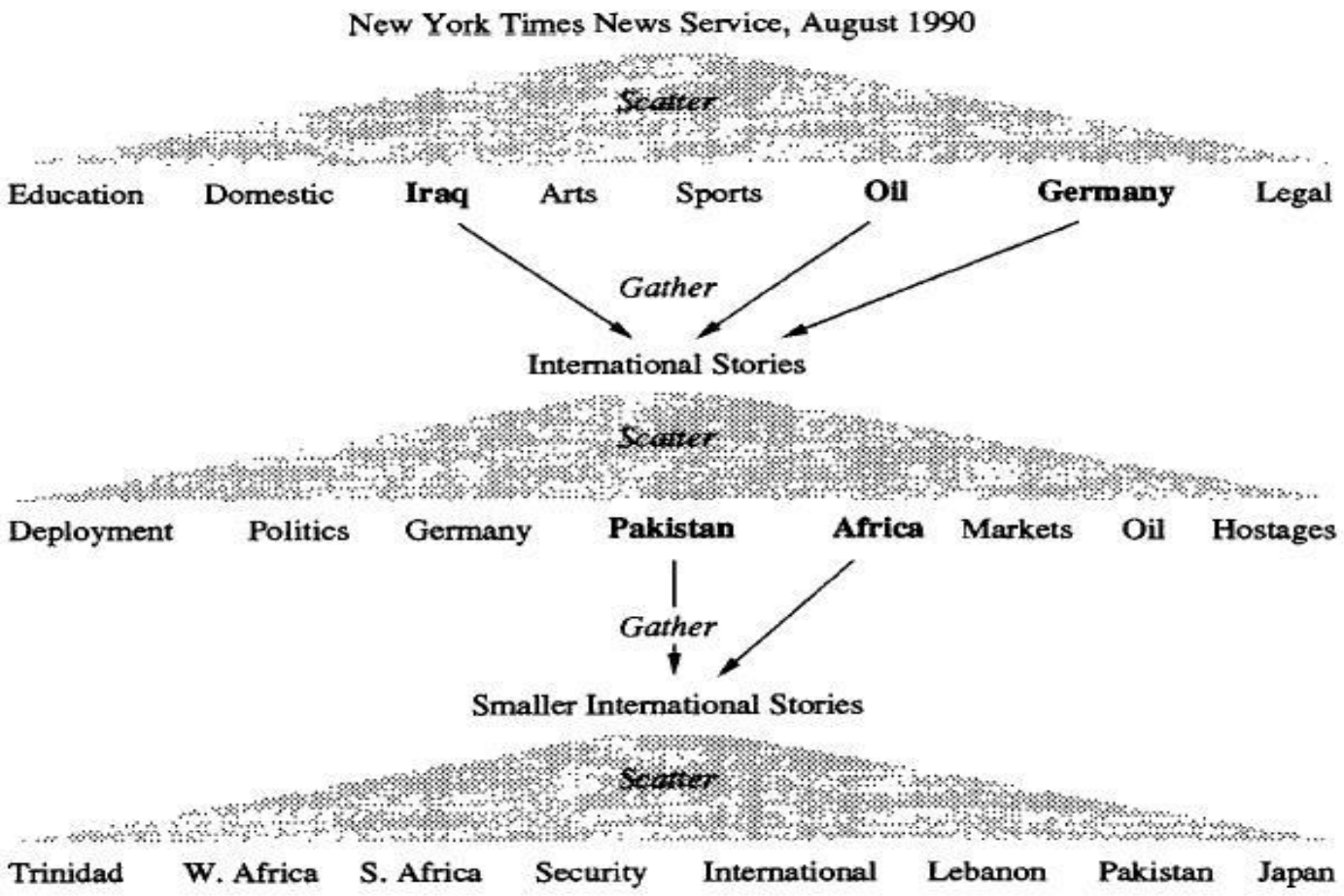


Figure 1: Illustration of Scatter/Gather

Application II

- Semi-novel
- Automatically generating term associations to aid in query expansion
  - Word mismatch
- Clustering
- Global / local analysis[7]

Global vs. Local Analysis

Global Analysis

- Expensive
- Clustering based on all documents
- Provides a thesaurus-like resource

Local Analysis

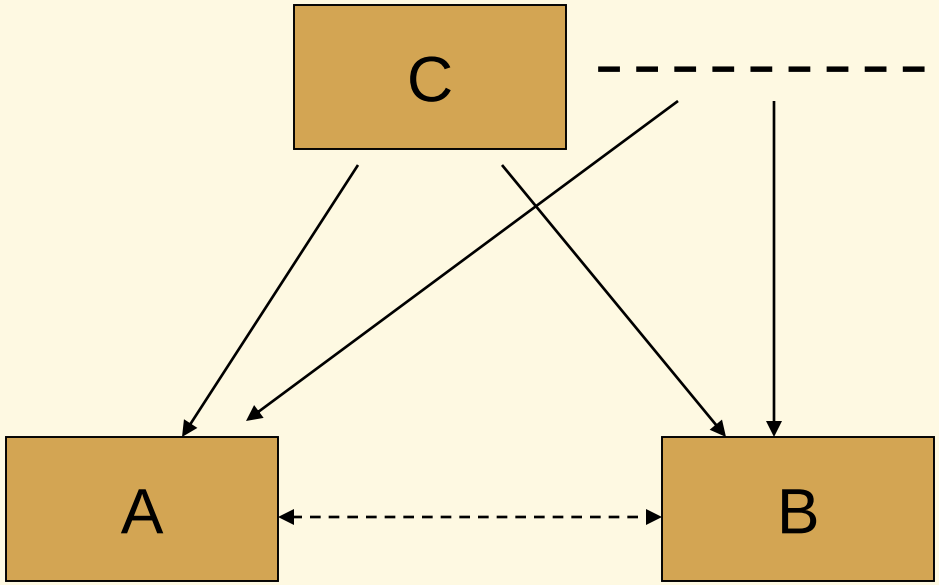
- Cost-efficient
- Clustering based on documents returned from previous query
- Only provides a small test collection

Application III

- Semi-novel
- Using co-citation analysis to find general topics within a collection or identify central web pages

Co-citation

- Bibliographic Co-Citation is a popular similarity measure used to establish a subject similarity between two items.
- E.g. basic idea of Google's algorithm



Co-citation Analysis Steps

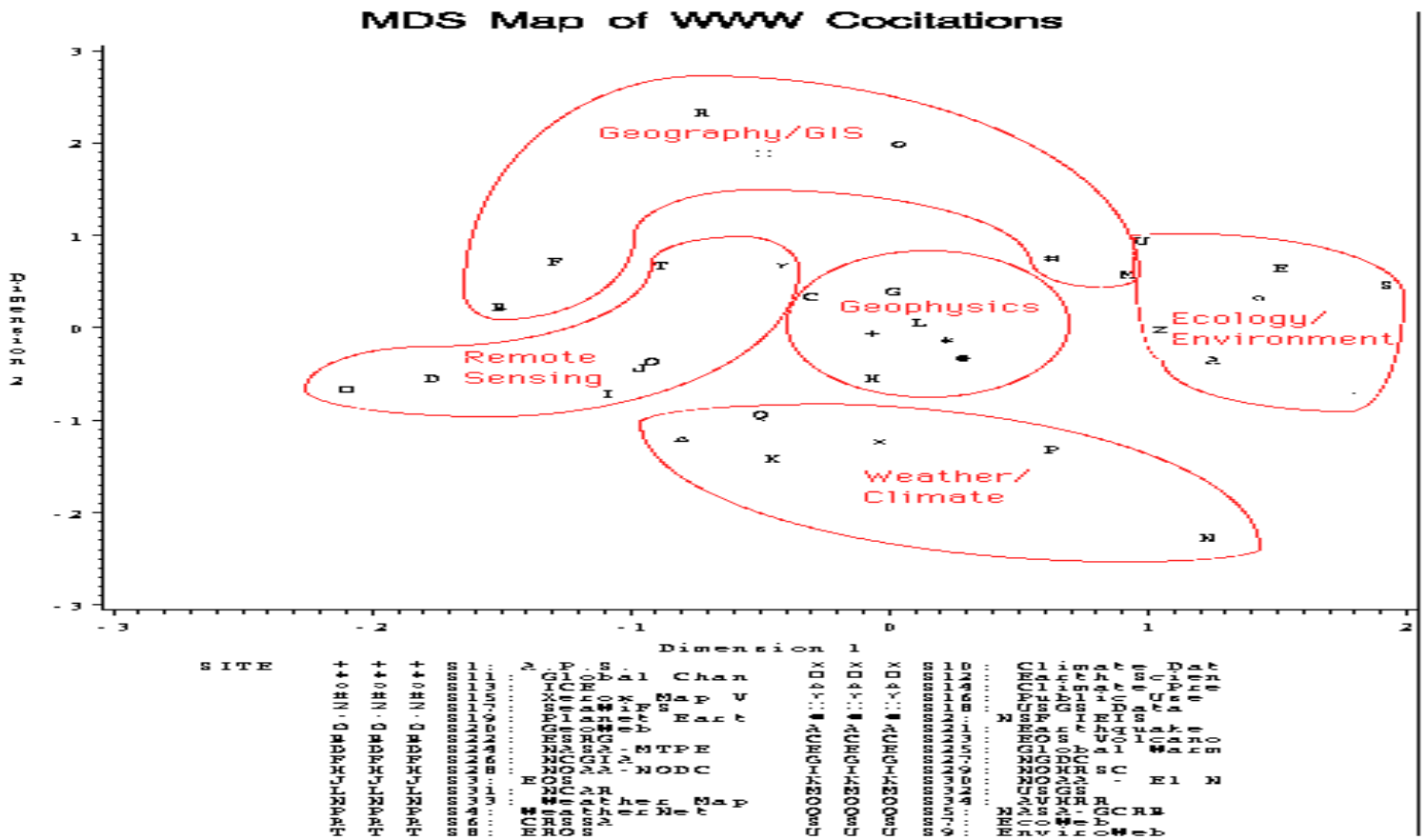
- Selection of the core set of items for the study.
- Retrieval of co-citation frequency information for the core set.
- Compilation of the raw co-citation frequency matrix.
- Correlation analysis to convert the raw frequencies into correlation coefficients.
- Multivariate analysis of the correlation matrix, using principle components analysis, cluster analysis or multidimensional scaling techniques.
- Interpretation of the resulting ``map'' and validation.

The Raw Co-citation Frequency Matrix

Raw Cocitation Frequency

ID																																	
	a1																																
a2	96	a2																															
a3	74	128	a3																														
a4	27	161	25	a4																													
a5	41	182	91	123	a5																												
a6	4	13	24	1	5	a6																											
a7	9	63	19	10	6	8	a7																										
a8	6	16	14	12	4	5	5	a8																									
a9	127	400	113	120	138	20	700	15	a9																								
a10	13	83	20	119	31	1	19	17	75	a10																							
a11	68	200	200	71	151	25	130	26	400	97	a11																						
a12	3	3	8	2	17	2	0	3	1	2	17	a12																					
a13	72	182	91	29	121	19	200	6	1000	23	300	8	a13																				
a14	63	185	82	106	109	6	18	11	109	90	173	3	85	a14																			
a15	132	181	47	164	25	40	58	26	600	50	154	5	200	36	a15																		
a16	5	20	35	15	17	20	33	6	44	16	92	8	45	8	48	a16																	
a17	75	180	150	168	110	14	119	32	700	165	200	24	184	85	300	106	a17																
a18	1	26	7	12	6	12	0	4	13	5	22	0	8	4	48	6	11	a18															
a19	27	13	5	66	2	5	8	3	86	4	11	0	16	45	75	5	51	2	a19														
a20	6	44	22	14	5	51	20	10	47	12	40	0	78	10	142	12	23	39	7	a20													
a21	20	53	37	101	15	2	95	3	200	38	43	1	42	81	170	20	200	3	65	19	a21												
a22	1	6	4	0	2	3	1	1	2	4	7	1	4	5	3	4	4	0	0	0	1	a22											
a23	11	69	76	12	15	16	31	21	86	17	75	2	45	23	91	33	200	3	6	23	63	3	a23										
a24	5	14	74	17	31	4	3	6	15	20	62	11	15	15	16	23	73	0	9	5	6	2	10	a24									
a25	5	38	19	6	11	5	128	0	167	27	124	1	130	35	42	31	65	1	1	3	23	2	29	4	a25								
a26	55	85	75	3	38	42	2	5	59	2	106	2	88	60	58	11	18	11	4	70	2	2	6	6	2	a26							
a27	48	131	59	39	31	33	58	9	159	31	152	2	146	46	200	52	123	29	5	51	35	5	59	20	34	37	a27						
a28	80	175	137	33	84	9	54	22	300	72	200	3	101	192	62	11	400	7	6	27	111	5	37	22	24	64	106	a28					
a29	49	112	81	56	66	16	7	10	66	26	116	2	68	117	32	13	28	6	2	23	4	2	9	10	6	58	53	106	a29				
a30	6	39	4	44	28	0	1	4	33	43	54	2	8	60	16	3	78	2	2	3	11	2	6	7	17	1	12	36	13	a30			
a31	136	300	148	185	128	20	50	23	300	80	300	6	200	187	400	48	194	29	8	32	43	5	56	30	47	92	300	200	120	67	a31		
a32	70	200	99	126	63	46	52	56	600	58	200	11	200	93	600	42	300	73	44	132	108	6	100	40	27	100	157	153	96	29	145	a32	
a33	6	40	2	117	24	2	2	2	26	10	12	0	11	18	43	4	21	1	10	4	16	0	10	3	3	3	4	5	4	9	27	21	a33
a34	88	151	152	25	128	17	13	24	109	59	200	16	95	119	115	28	157	16	6	27	14	2	70	52	18	105	86	127	86	37	175	81	9

Result


MDS Map of WWW Cocitations
Geography/GIS
Geophysics
Ecology/
Environment
Remote
Sensing
Weather/
Climate
Dimension 1
Dimension 2

Artificial Intelligence

- **Artificial intelligence (AI)** is a branch of computer science and engineering that deals with intelligent behavior, learning, and adaptation in machines.

Self-Organizing Map (SOM)

- One category of neural network models.
- Neighboring cells in a neural network compete in their activities by means of mutual lateral interactions, and develop adaptively into specific detectors of different signal patterns.
- Here, learning is called competitive, unsupervised, or self-organizing.[6]

Self-Organizing Map (SOM)

- An example of how SOM works?
  - Each node has two vectors: input vector, weight vector (location)
  - Input vector (Red, Green, Blue)
  - Red (255, 0, 0) Green (0, 255,0) Blue (0, 0, 255)
  - Randomize the weight vectors of nodes in the map
  - Calculate the Euclidean distance formula to find out the smallest difference between the input vector and the weight vector (Best Matching Unit or Winner)
  - Pulling neighbors closer to the input vector
  - Repeat a large number of cycle.

Self-Organizing Map (SOM)

- In attempting to devise neural network models for linguistic representation, the difficulty is how to find metric distance relations between symbolic items.

SOM Application

Web Anlysis

- Problem: directory-based search engines such as Yahoo! analyze, index and categorize web content manually.
- Solution:
  - high-precision noun phrase indexing was performed to each page
  - A vector space model of noun phrases and their associated weights were used to present each page
  - All pages were categorized by a SOM clustering program[4]

Folksonomy

- **folksonomy** is an Internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels that categorize content such as Web pages, online photographs, and Web links.

Folksonomy

- Benefits:
  - Lower content categorization costs
  - respond quickly to changes and innovations
  - the capacity of its tags to describe the "aboutness" of an Internet resource

Folksonomy

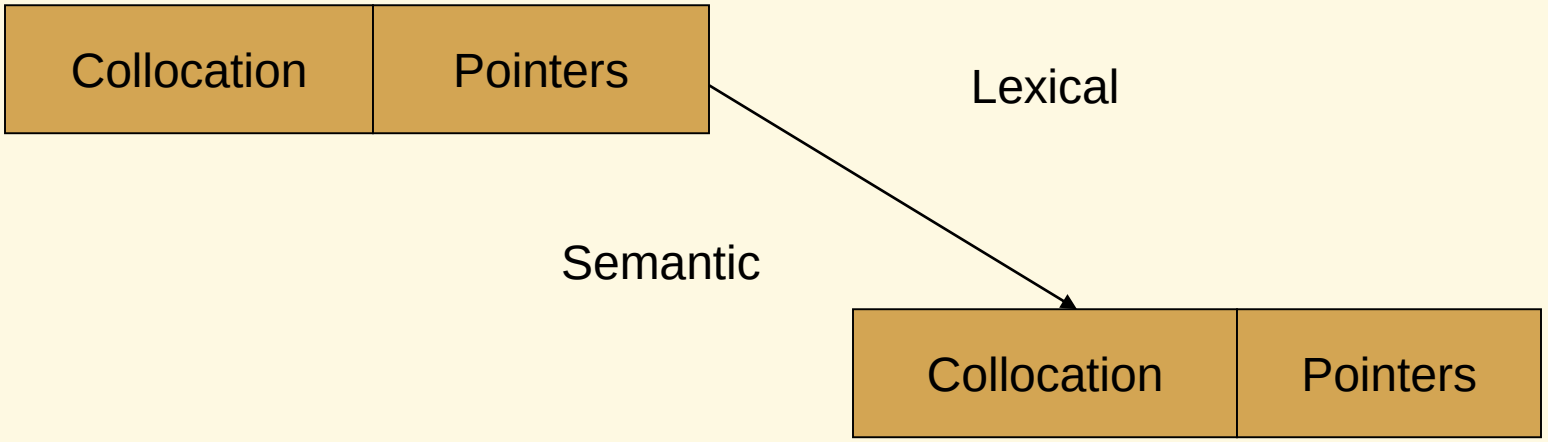
- Lack of standard: polysemy, synonym
- Meta noise: inaccurate or irrelevant metadata

Natural Language Processing

- **Natural language processing (NLP)** is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages.
- Statistical natural language processing uses stochastic, probabilistic and statistical methods to resolve some of the difficulties : e.g. text segmentation, word sense disambiguation

Application I

- Semi-novel
- WordNet
  http://wordnet.princeton.edu/
- Structure:



WordNet

- Nouns and verbs are organized into hierarchies based on the hypernymy or hyponymy.
- Adjectives are arranged in clusters containing head synsets and satellite synsets:
  - Head synsets: clusters of antonymous pairs/triplets
  - Satellite synsets: a concept similar in meaning to that of the head synset

WordNet

- A synset for a pertainyms contains only one word or collocation and a lexical pointer to the noun that the adjective is "pertaining to".
- A synset for an adverb contains a lexical pointer to the adjective from which it is derived.

Glossary

- Hypernymy: **Y** is a hypernym of **X** if **X** is a (kind of) **Y**
- Hyponymy: **X** is a hyponym of **Y** if **X** is a (kind of) **Y**

Application II

- Novel
- Bioscience: deduce hypotheses (National Centre for Text Mining in UK)

An Example

- When investigate causes of migraine headaches, we extracted various pieces of evidence from titles of articles in the biomedical literature:
  - 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 levels of magnesium inhibit SCD
  - migraine patients have high platelet aggregability
  - magnesium can suppress platelet aggregability

Application III

- Novel
- Using text to uncover social impact
  - the effects of publicly financed research on industrial advances
  - 1987 ~ 1988, 1993 ~ 1994; 397,660 patents issued
  - found 242,000 identifiable science references and zeroed in on those published in the preceding 11 years (80%)
  - 109,000 of these references to known journals and authors' addresses in computer database
  - a core collection of 45,000 papers, after eliminating redundant citations to the same paper, and articles with no known American author
  - look up the papers and examine their closing lines, which often say who financed the research

Reference

- [1] Ronen Feldman and Ido Dagan. "Knowledge discovery in textual databases (KDT)", *Knowledge Discovery and Data Mining* (1995), 112-117.
- [2] M. Hearst. "Untangling text data mining". *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
- [3] chen 2002

Reference

- [4] Hsinchun Chen, "Knowledge Management Systems: A Text Mining Perspective", *Knowledge Computing Corporation*, 2001.
- [5]Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. "Scatter/Gather: A cluster-based approach to browsing large document collections". *In Proceedings of the 15th Annual International ACM/SIGIR Conference1992*, pages 318-329, Copenhagen, Denmark.

Reference

- [6]Teuvo Kohonen. “The Self-Organizing Map”. Proceedings of the IEEE (1990), 78, 9, 1464-1480.
- [7] J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4-11, Zurich.

Reference

- [8] Ray R. Larson. Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *ASIS '96: Proceedings of the 1996 Annual ASIS Meeting*.

Thank You !

Munawar, PhD