Master of Technology in Knowledge Engineering

Text Mining

Association and Trend Analysis

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
email: zhenzhen@nus.edu.sg

Agenda

- Analyzing text over time
- Association/link analysis in general
- Text link analysis

What's next after getting concepts?

- With the extracted concepts available, you can already do your analytics on them. For example, assuming you are working on customer reviews data:
  - You can do a document count on the concepts you are interested in, e.g. "expensive" (How many customers are complaining about the pricing of your products/services?)
  - If you have rating score available for each review, you can test whether complaining "expensive" affects the customer's rating.
  - If you have time information for each review, you can plot and see the number of such records per week/month/quarter (Is the number of such customers increasing or decreasing?)
  - Association of "expensive" and other concepts representing details.
  - Etc.

Analyzing Text over Time

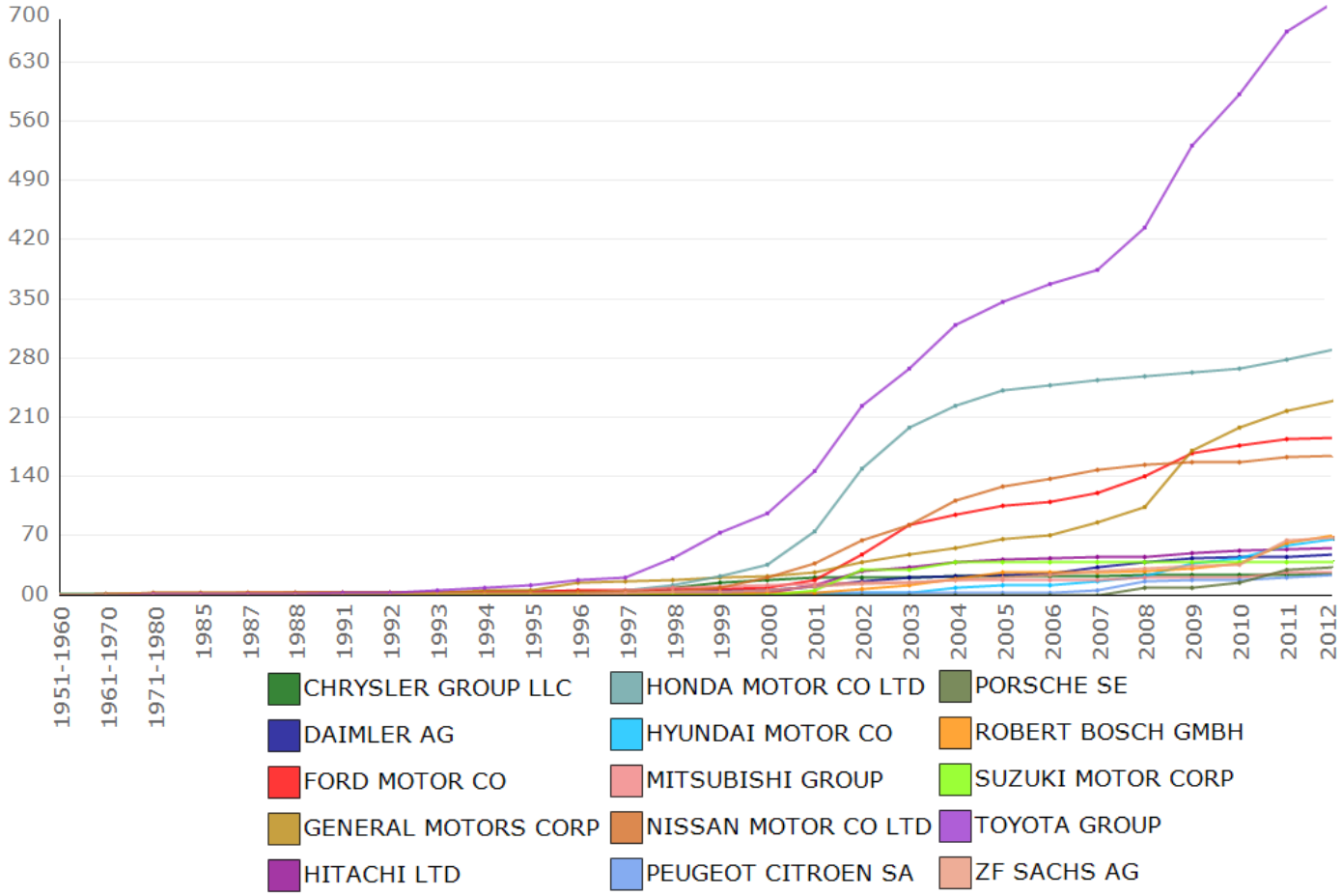
- Early text mining systems tended to view a document collection as a single, monolithic entity – a static set of textual documents.
- Many applications benefit from viewing the collection in terms of subsets or divisions defined by the date and time stamps.
- This view allows a user to analyze similarities and differences between concept relationships across the subsets in a way that better accounts for the change of concept relationships over time.
- => **Trend Analysis**

Trend Analysis

- The analysis of concept distribution behavior across multiple document subsets over time, e.g. collections from the same news feed or social media but from different points in time
- It can better account for the evolving nature of concept distributions and relationships in a collection.
  - E.g. compare the distribution of topics within financial news from the first quarter to those from the second quarter, etc.
  - Highlight the topics whose proportion changed between two time points
  - Highlight to the user specific trends (like upward trend) across different time points

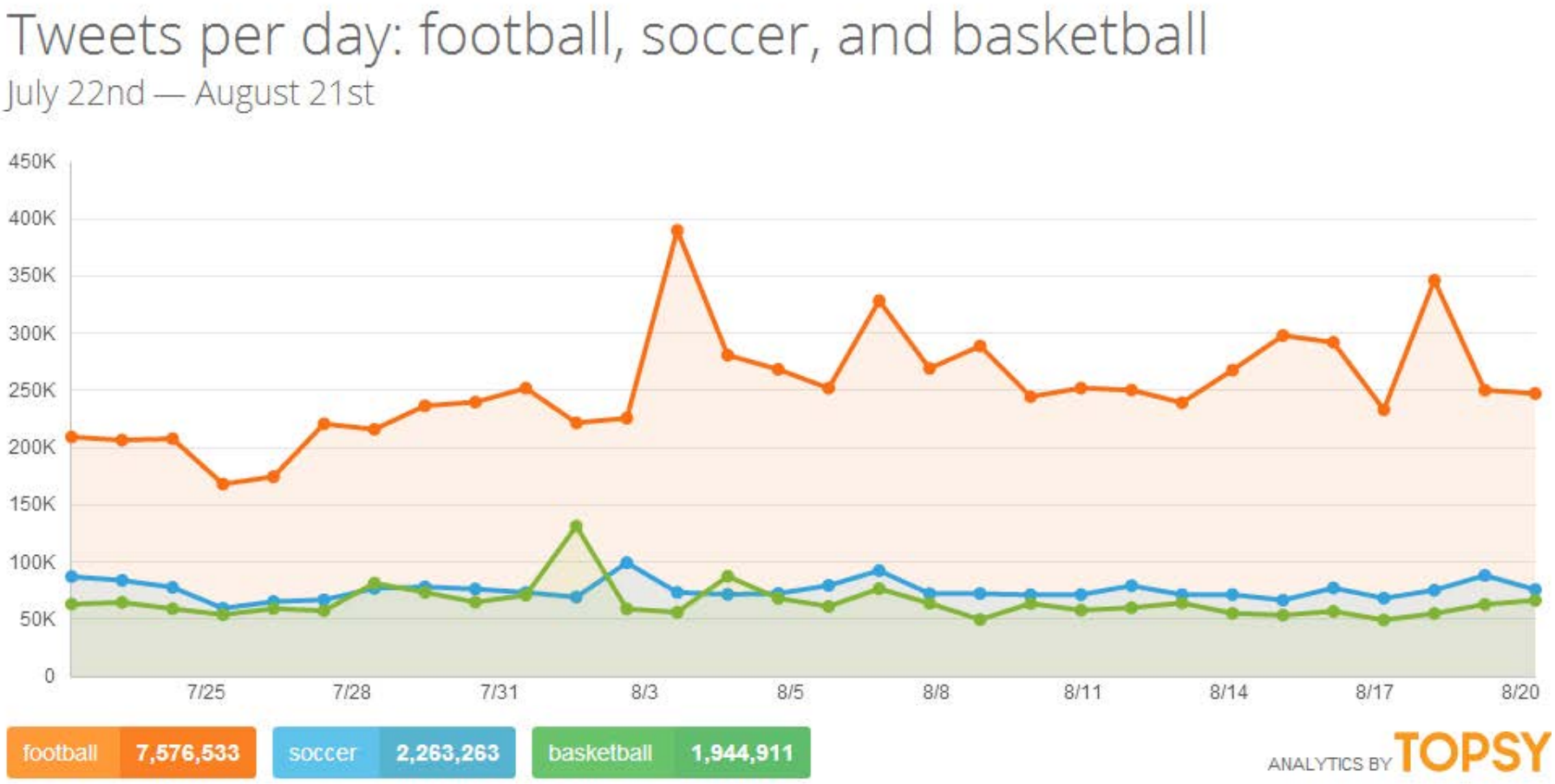
Example: Patent Assignee Trends

- From Patent iNSIGHT Pro



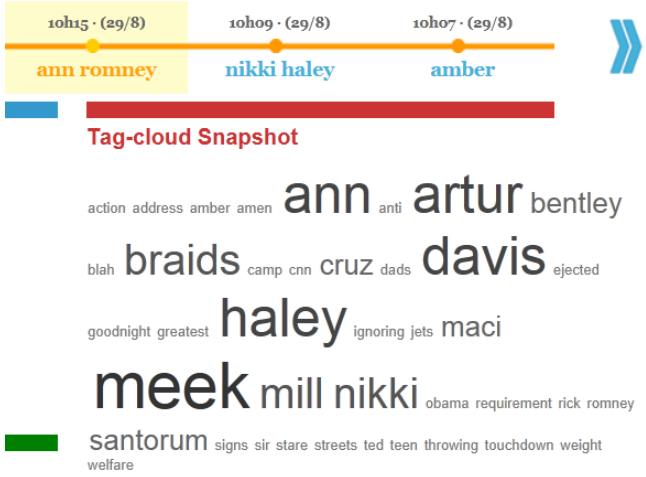
http://www.patentinsightpro.com/techreports.html

Example: Trends on Twitter



Another Example: Trends on Twitter

- And nice animation of wordcloud showing the changing hot topics



Analyzing Text Streams

- Monitor text streams online to dynamically detect and track topics, detect new events
  - Discover the themes/topics in text
  - Create an evolution graph of themes
  - Study the lifecycle of themes
- Considered very hard and still in active research, especially in the context of social media
  - Some methods proposed, for example, process documents sequentially to determine if an incoming document corresponds to a new event, by comparing it with the last m documents.
  - Often use weights based on the recency of a document
  - Common approaches to find broad topics include document clustering, latent topic modeling, etc.

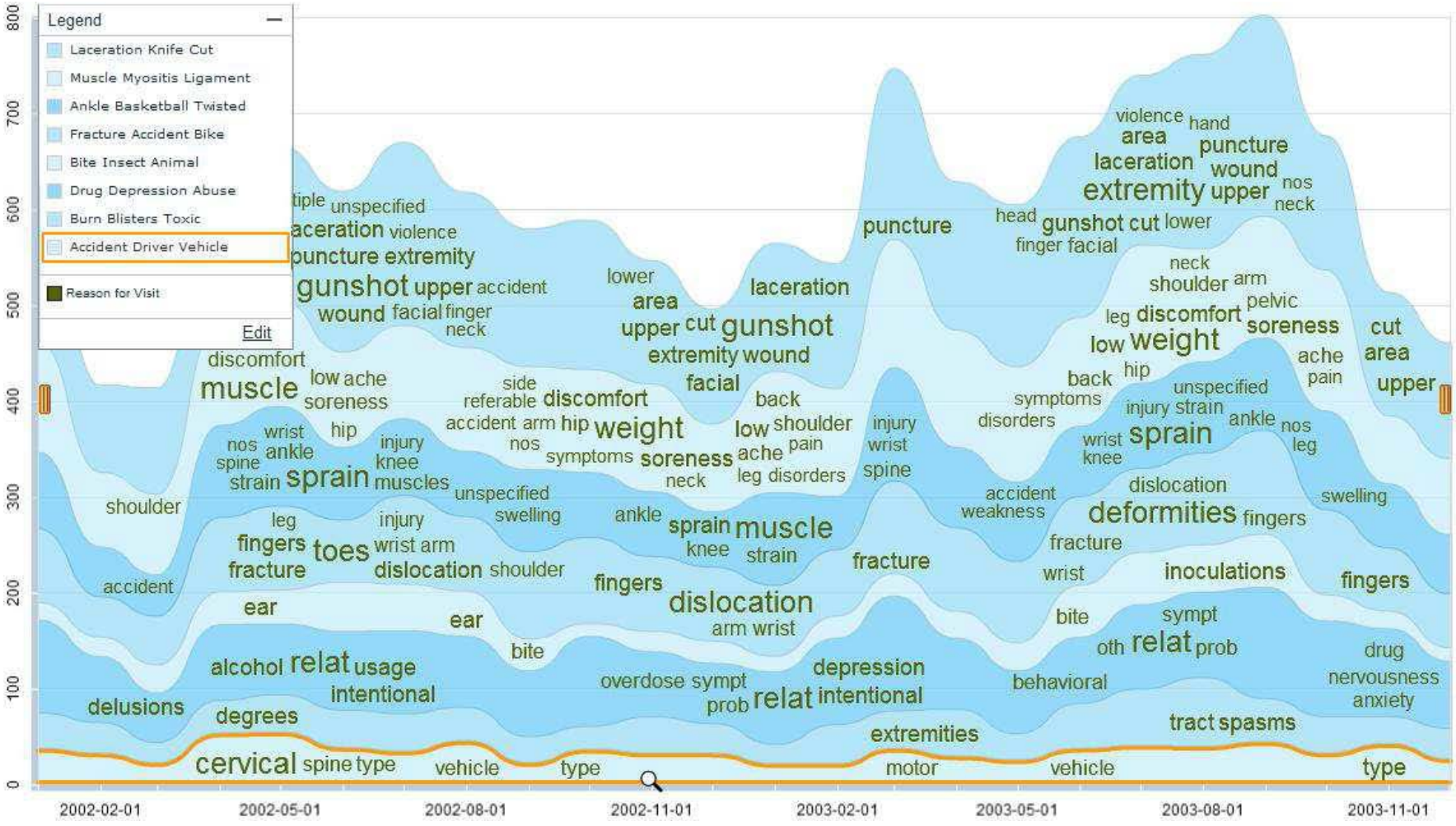
Microsoft TIARA

- Text Insight via Automated Responsive Analytics - a visual exploratory text analytic system
- Provide a time-sensitive topic-based summary
- Help answering questions like:
  - What are the major topics in the documents?
  - What are the most active topics over the last three months?
  - What are the key concepts mentioned in the aforesaid topics?
  - How have the most active topics evolved over time?
- Applied to email summarization, patient record analysis, etc.

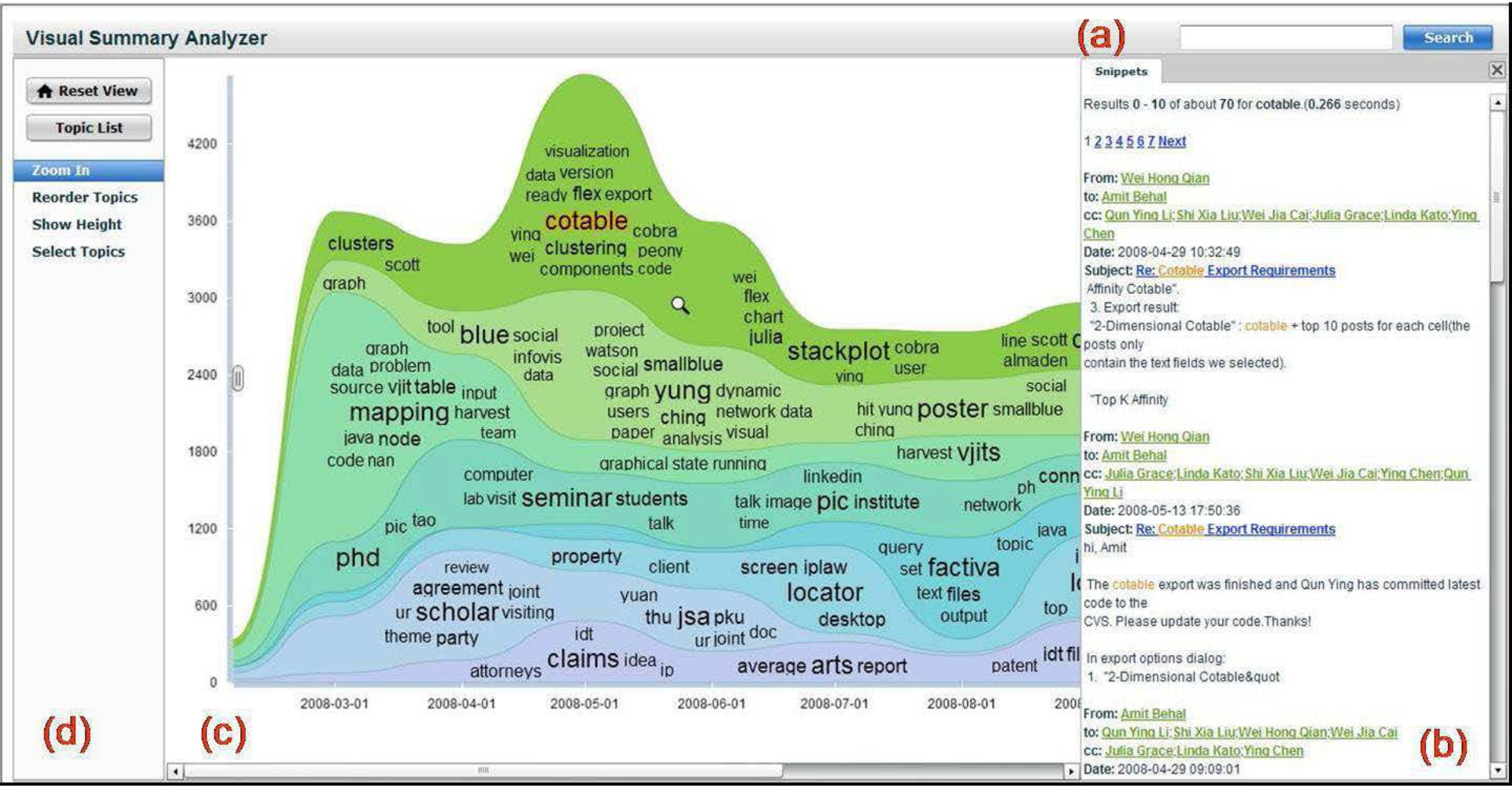
The Working of TIARA

- Given a collection of documents, topic analysis techniques (e.g. LDA) are first applied to summarize the docs into a set of topics.
  - A topic represents the thematic content common to a set of text documents.
  - Each topic is characterized by a distribution over a set of keywords.
  - Each keyword has a probability measuring the likelihood of this keyword appearing in the related topics.
- Time-sensitive keywords are derived to depict the content evolution of each topic over time.
- Interactive text visualization techniques are used to explain the topic-based summarization results.

TIARA: Summary of "Reason for Visit"



TIARA: Evolution of Email Topics



Association: Co-occurring Words

- Words appearing together in the same document

- Words appearing consecutively together, or in close vicinity in a defined relation, in the same document (expanded definition of *collocations*)

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.

Association/Link Analysis in Data Mining

- In Data Mining, Link Analysis is traditionally used in Market Basket problems
  - To determine what products are purchased together or likely to be purchased by the same person
  - Commonly applied to identify cross-sell and up-sell opportunities
    - E.g. Items purchased on a credit card (e.g. rental cars, hotel rooms) give insight into the next product the customer may buy
  - Associations between products are visualized or represented as ***association rules*** generated from ***frequent item sets*** discovered from transaction data

Data for Link Analysis

- A list of transactions, for example, at a convenience store, each including a list of items/products

Transaction Data

No.	Products purchased
1	Frozen pizza, cola, milk
2	Milk, potato chips
3	Cola, frozen pizza
4	Milk, peanuts
5	Cola, peanuts
6	Cola, potato chips, peanuts
…	

Tabular Data

No.	Pizza	Cola	Milk	Chips	Peanuts
1	T	T	T	F	F
2	F	F	T	T	F
3	T	T	F	F	F
4	F	F	T	F	T
5	F	T	F	F	T
6	F	T	F	T	T
…					

Does this look familiar?

Link Analysis for Textual Data

- Applied to textual data after pre-processing stage:
  - Instead of **transactions**, we have **documents** (in vector model or TDM).
  - Instead of **products/items**, we have **terms/concepts or other derived features**.
- Analysis is based on Frequent *Term/Concept* Set discovered from documents:
  - a set of terms/concepts or other derived features represented in the document collection with co-occurrences ≥ minimal *document support* level, i.e. appearing together in at least *s* documents
  - Apriori-like algorithms can be used to discover the frequent term/concept set

Problem in Generating Frequent Term Sets

- Found by counting the *document support* of possible combinations of terms, which can be combinatorially explosive

E.g. If 100 terms are in your TDM

No. of terms in term set	No. of combinations
1	100
2	4,950
3	161,700
4	3,921,255
5	75,287,520
6	1,192,052,400
8	186,087,894,300

How to Derive Frequent Term Set

Using Apriori algorithm

Scan	Candidates	Large item sets
1	{hotel}{expensive}{service}{excellent}{location}{stay}	{hotel}{expensive}{service}{excellent}{location}
2	{hotel, expensive} {hotel, service} {hotel, excellent} {hotel, location} {expensive, service} {expensive, excellent} expensive, location} {service, excellent} {service, location} {excellent location}	{hotel, expensive} {service, excellent}
3	…	…

Only consider term sets with size ≥ N

Apriori Algorithm

Reduces the number of sets to consider:

1. Count the support of every individual item from the transaction data, find those exceeding the minimum support, and output the frequent/large 1-item sets.
2. Take the large item sets from the previous step, generate 2-item sets as candidate item sets, count the support for these candidates, and determine large 2-item sets meeting the minimum support.
3. Large 3-item sets are generated similarly based on large 2-item sets.
4. The process continues until no new large item sets are found.

Association Rules

- Association rules can be generated from the discovered frequent term/concept sets.
- In MBA, the association rules are useful for spotting cross-selling opportunities, e.g.

- In Text Analytics, such rules are probably more meaningful for association between terms/concepts and other features, e.g.

If a customer mentions "expensive"

then he will not recommend it to others

Association Rule Generation

- Straight-forward algorithm (in Apriori): generate all possible rules from a large item set, and select those with evaluation score above a threshold
  - E.g. If there are three items A, B, C, then possible rules are

    If A and B then C

    If A and C then B

    If B and C then A

- Typical score :
  - **Support** – the number or percent of the documents containing all the concepts in the given rule, i.e. the co-occurrence frequency
  - **Confidence** – the percentage of documents that include all the concepts in RHS with the subset of those documents containing all the concepts in LHS, i.e. the percentage of the time that the rule is true
- The minimum support and minimum confidence threshold of the rules can be defined by user.

Association Rule Evaluation

If "expensive" then "will not recommend" (R1)
If "heavy" then "will not recommend" (R2)

"expensive" & "will not recommend" (20)
"heavy" & "will not recommend" (20)
"expensive" (40)
"heavy" (25)

"will not recommend "(50)
"will recommend (50)
Total (100)

- **Support (**probability of getting that combination)
  - R1 support = 20% (20 out of 100 included "expensive" & "will not recommend")
  - R2 support = 20% ((20 out of 100 included "heavy" & "will not recommend")
- **Confidence (**Support combination/ Support condition (LHS))
  - R1 confidence = 50% (20 out of 40)
  - R2 confidence = 80% (20 out of 25)

Detecting Collocation

- The collocations can be discovered by comparing the number of times two or more words appear together with the number of times they appear in other contexts.
- The challenge lies in separating those from words that randomly appear together.
- Common strategies:
  - Statistical hypothesis testing
  - Information theoretic analysis
  - Combination of a part of speech tagger and a simple frequency filter

Collocations

- In linguistics, a collocation is an expression consisting of two or more words that correspond to some conventional way of saying things.
- Examples:
  - Strong tea
  - Weapons of mass destruction
  - Broad daylight
  - Make up
  - Kick the bucket
- Limited compositionality, with added meaning

Collocation Detection – Frequency-based

- Combination of a part of speech tagger and a simple frequency filter – simple but effective!
  - Based on linguistic patterns of collocations, usually as combinations of nouns and adjectives
  - A part-of-speech tagger is required to find POS for each word, and identify phrases following such patterns.
  - Then apply sorting based on frequency to identify collocations
  - Examples of POS patterns

Adjective + noun	*third quarter*
Adjective + adjective + noun	*executive vice president*
Noun + preposition + noun	*earnings per share*

Examples from New York Times News

- Using raw frequency vs. with POS filtering

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N

Collocation Detection – Statistical hypothesis testing

- To detect those that occur together more often than chance
- Formulate a ***null hypothesis H_0*** (no association beyond chance) and calculate the probability that a pair of words would co-occur if H_0 were true, and then ***reject H_0*** if p is too low ($p<0.05$, 0.01, 0.005, or 0.001)
- The t-test, the chi-square test, and the likelihood ratio
- Based on contingency table where the cells of the table contain the counts of times each appears both in isolation and together.

Is "new companies" a collocation?

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

- The chi-square distribution indicates that at probability level of =0.05, the critical value of chi-square is 3.841.
- We can not reject the null hypothesis.

Collocation Detection – Mutual Information

- Based on the principles of information theory
- If the appearance of the first word provides a strong indication that the second word will follow, then those two words are marked as a collocation.

$$
\begin{aligned}
I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\
&= \log_2 \frac{P(x'|y')}{P(x')} \\
&= \log_2 \frac{P(y'|x')}{P(y')}
\end{aligned}
$$

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

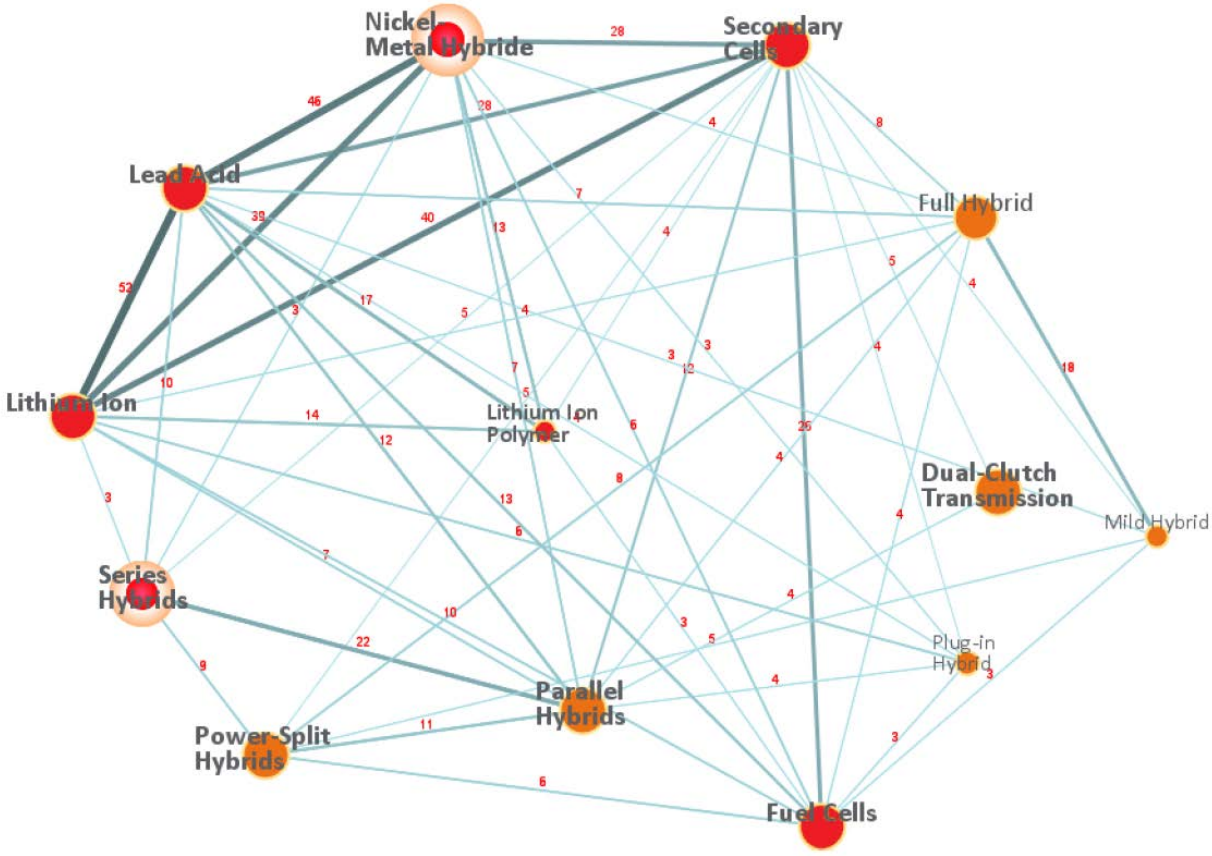
$$
I(\textit{Ayatollah}, \textit{Ruhollah}) = \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \approx 18.38
$$

Visualizing Links

- Huge numbers of links can be discovered between terms and concepts, therefore it's helpful to have ways to visualize and interactively explore the results
- View links as graph
  - Nodes: terms or concepts or other extracted features
  - Links between nodes: connecting associated nodes
  - Thickness of the line – document support
  - Different icons can be used to indicate the type of features
- Typically allow filtering of the displayed nodes

Example: Patent Analysis

- Analysis of hybrid car patents from Patent iNSIGHT Pro showing links between battery types and transmission modes



http://www.patentinsightpro.com/techreports.html

Focus Analysis between Features of Interest

- In practice, especially with long documents, such link analysis often generates a large number of associations that are hard to interpret.
  - You can restrict your analysis to the associations between concepts/features you are interested in, e.g. products and negative terms.
  - You can also adjust the threshold value for support/confidence
- Remember: Association found through pure within-document co-occurrence may be misleading.

  The storyline is great…

  …with such a terrible actor*…*

  => Restrict that only words in defined association patterns are counted (e.g. adj + n)

Example from SPSS Modeler

- A web graph showing the association patterns

References

- Agrawal and Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference of Very Large Data Bases (VLDB)*, Santiago, Chile, 1994.
- Feldman and Sanger. *The Text Mining Handbook: Advanced approaches in Analyzing unstructured Data*, Cambridge University Press, 2006
- Wei, Furu, et al. Tiara: a visual exploratory text analytic system. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.
- Manning and Schutz. Chapter 5 Collocation. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.