Tutorial on

Web and Text Mining for opinion/trend analysis

IEEE Conference of WI and IAT, 2007

Dr. Lipika Dey
Tata Consultancy Services
Delhi, India

Overview of tutorial

- Why mine the web?
- What to mine?
  - User history, document links, contents
- How to mine?
  - Tools and Techniques

Organization of talk

- Introduction
- Web Usage Mining
- Web Structure Mining
- Text Mining
- Opinion extraction
- Trend analysis

The Web - A Snapshot

Why mine the web?

- The Web tells
  - What the world is looking for
  - What the world is thinking
- The Web determines
  - What the world will look for
  - What the world will look like

The Web in control

- What are users' reaction to a news?
- What are user opinions about a product?
  - *Shape the future*
- How do comparative products fare?
- How is a peer work rated?
  - *Influences my decision*

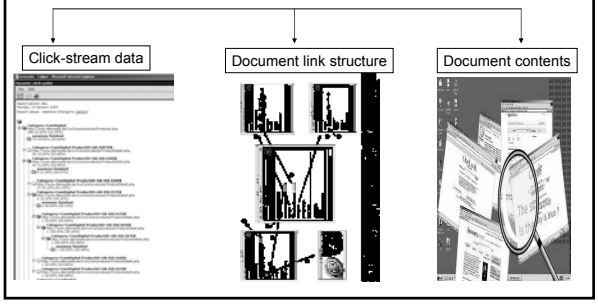
Mining versus Search

- Search refers to retrieving relevant information from web
  - **Goal oriented**
- Mining refers to discovery of useful, hitherto unknown information from web usage data and web contents
  - **Opportunistic**

Mining the web

- Discovery of useful **patterns** from the web
- Can provide hitherto unknown information
  - Web usage
  - Web structure
  - Web content

What is mined?



Mining user access data

- Analyzes large volume of data generated by servers
  - To study user behavior
  - Content usage statistics
- Useful in predicting future
  - Uses Data Mining techniques

Web structure analysis

- Analyze the nodes and links of web
- Information connectivity
  - Social link analysis
  - Search Technology
- Graph-theoretic techniques

Web content mining

- Discovering information from unstructured content
- Knowledge oriented activity
  - Analyzes intent, opinion, views, reviews
- Goes beyond retrieval to **analyze** content and **infer** new information

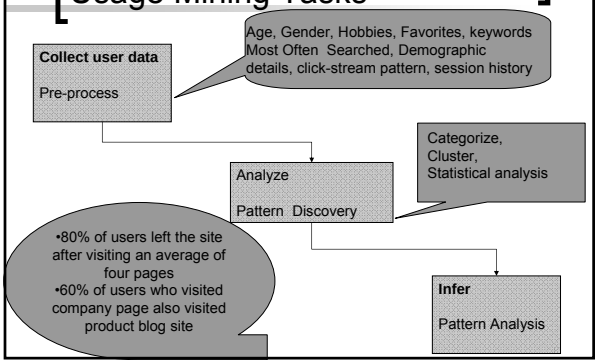
Web Usage Mining

Data generated due to web usage

Web usage mining

- *Collective analysis reflects user opinion*
- *Mine user profiles of visitors*
  - Provides insight into adequacy of information regarding user needs
  - Plan marketing strategies and promotional campaigns
- *Extract group behavioral patterns*
  - *Recommend similar online contents and products*
- *Recommendation plays a big role in influencing opinion*

Usage Mining Tasks



Sample trends - from query analysis

- 1) Almost all of the ten biggest US markets for MySpace are on the West Coast
- 2) There's more interest in Bluetooth than in Wifi
- 3) PR is starting to come close to catching advertising. AdAge still bests PR Week
- 4) David Hasselhoff's popularity in Germany is declining
- 5) Jerry Lewis's popularity in France is rising
- 6) Blackberry is widening its lead over the Treo
- 7) Zacharias Moussaoui finally beat OJ Simpson in searches
- 8) Hockey is starting to surpass baseball in popularity, but they fall way behind football (no matter which way it is defined)
- 9) New York still tops LA!
- 10) The Kryptonite Lock got more PR from the blog blow-up than any other event during the past two years

Recommender systems

- Backbone of modern marketing
- Predict items that may be of interest to user
  - Provide a list of top ten popular albums
  - Provide a list of new"Jazz" albums if user has searched for "Jazz" most often in the past

Techniques

- Association Mining
  - Frequently occurring patterns
- User groups
  - K-nearest neighbors with feature vectors
- Sequential Pattern Mining
  - Click-stream analysis
- Recommend items
  - Weighted average of ratings

Handling symbolic objects

- User profiles are symbolic objects
- Distance functions for defining neighbours N
  - $D(a,b) = d_{gender}(a,b) + d_{age}(a,b) + \ldots$
  - $d_{gender}(a,b) = [1,0]$
  - $D_{age}(a,b) = |a.age - b.age|/max$
- Weighted average of neighbour's ratings
  - $P(a) = \Sigma_{u \epsilon N} r(u) / (\Sigma_{u \epsilon N} 1/D(a,u))$

Content based rating

- Computing similarity of two contents i and j belonging to Itemset I
  - $S_{i,j} = \text{Cos } \theta = (V_i.V_j)/|V_i||V_j|$
- Predicted user rating on similar item
- $P(a,j) = \Sigma_{i \epsilon I} r(i).S_{i,j} / \Sigma_{i \epsilon I} S_{i,j}$

Probabilistic Model

- Probability of a rating value given known ratings
  - $Prob(r_{a,j} = k \mid r_{a,j,}\ i \epsilon I_a)$
  - $i \epsilon I_a$: Set of items rated by user a
- Probability can be learnt using Bayesian Network learning
- Expected rating
  - $P_{a,j} = E(r_{a,j}) = \Sigma_{k \epsilon R} k.\ Prob(r_{a,j} = k \mid r_{a,j,}\ i \epsilon I_a)$

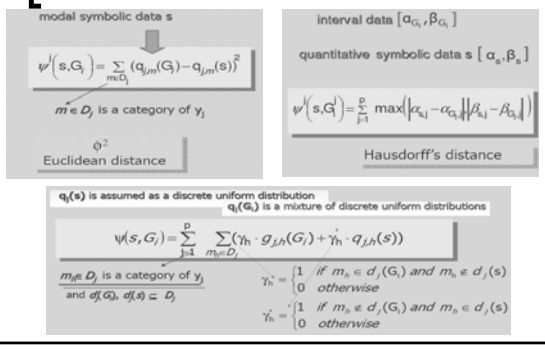
Other models

- Modal variables
- Obtained as functions of distributions of objects in a cluster are used to represent the clusters

Dynamic Clustering algorithm

- Dynamic clustering algorithm optimizes a criterion Δ which achieves the best fitting partition P* for a collection of objects O into k clusters
- $P^* = \{C_1, C_2, .., C_k\}$.
- Clusters are represented by representative elements $\{G_1, G_2, \ldots, G_k\}$

Sample Distance function [Lechevallier et al., 2006]



Mining the links

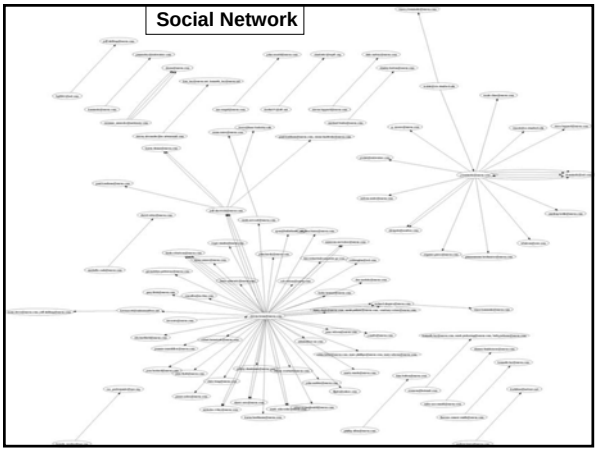
Links are exploited for search

Web Structure Mining

- Generate structural summary about web sites and web pages
- Discovery of interesting and informative facts about connectivity from web tuples in a web table
  - Frequency of local links
  - Frequency of links that are interior and the links that are within the same document
  - Frequency of links that are global and the links that span different Web sites
  - Information measuring the frequency of identical entries in a table or collection of tables.

Link Analysis for social networks

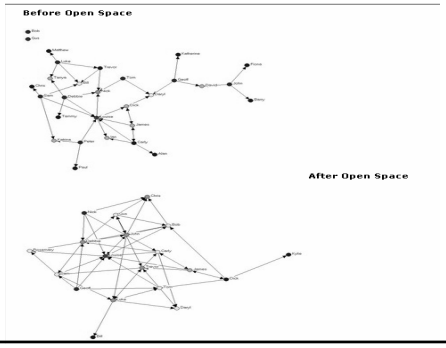
- Given any two persons - find a path between them that has less than 6 edges.
- Eg. Kevin Costner links to Kevin Bacon by using one direct link: Both were in JFK.
- Julia Louis-Dreyfus of TV's Seinfeld, however, needs two links to make a path: Julia Louis-Dreyfus was in Christmas Vacation (1989) with Keith MacKechnie. Keith MacKechnie was in We Married Margo (2000) with Kevin Bacon.



Impact of social network analysis

- Marketing strategies
- Influence of one community over another analyzed
- Analysis of Enron Email corpus
  - Who was passing information to whom?

Reach increases



Mining Contents

Text Mining
Analyzing opinionated text
Understanding opinion orientation
Inferring trends

Text Mining

- *Discovers new pieces of knowledge*
- A key element is the *linking* together of the extracted information together to form *new facts* or *new hypotheses* to be explored further by more conventional means of experimentation.
- *Find overall trends in textual data*

Text Mining is not Information Extraction

- Information Extraction – yields informative chunks of data extracted from text
- Uses Natural Language Processing Techniques
- IR can answer user posed queries
- Does not address the issue of establishing **"interestingness"** of information

Text Mining applications - commerce

- Find "important" business houses from web houses
- Frequently occurring names may not be "*interesting*" since well known
- Really interesting – identify names that are worth keeping a watch on
- *How???????*
  - *– analyze views, reviews, news, ……*

Potential in Bio Sciences

- Swanson et al. (1987 – 1997)
- Causes of Migraine headache
  - 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
- ***Magnesium deficiency may play a role in some kinds of migraine headache***

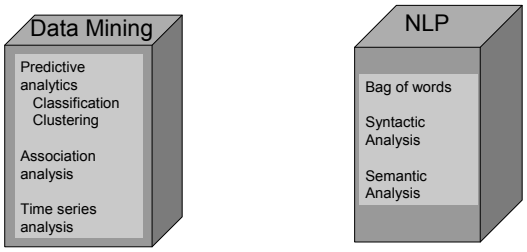
Social Impact [Narin et al.1997]

- ***Technology industry relies more heavily than ever on government-sponsored research results.***
- Examined the science references on the front pages of American patents
- 73.3 percent had been written at public institutions - universities, government labs and other public agencies, both in the United States and abroad.

Document Types on web

- Structured documents
  - Output from CGI
- Semi-structured documents
  - Seminar announcements
  - Job listings
  - Ads
- Free format documents
  - News
  - Blogs
  - Reviews
  - Scientific papers

Tools and Techniques



Text Representations

- Character Trigrams
- Words
- Linguistic Phrases
- Non-consecutive phrases
- Frames
- Scripts
- Role annotation
- Parse trees

Unique aspects of Text Mining

- **Feature extraction.**
- **Very large number of features that represent each of the documents.**
- **The need for background knowledge.**
- **Even patterns supported by small number of document may be significant.**
- **Huge number of patterns, hence need for visualization, interactive exploration.**

Text mining process

- Text preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Statistics
- Text/Data Mining
  - Classification- Supervised learning
  - Clustering- Unsupervised learning
- Analyzing results

Text Mining Tasks

- **Given:**
  - A source of textual documents
- **Find:**
  - Sentences with relevant information
  - Extract the relevant information and ignore non-relevant information (important!)
  - **Link** related information and output in a predetermined format

Opinion mining

- Refers to identifying user opinions about various objects on web
- Available in the form of
  - Explicit User responses
  - Implicit responses
  - Text documents

Why Opinion Mining

- People are concerned about opinions
- Opinions of different polarities in documents are useful references or feedbacks for governments or companies helping them improve their services or products
- Three categories of opinions
  - positive, neutral and negative

Defining Opinion Extraction

- Opinion sentence extraction
- Opinions are usually about a theme
  - Target and its features to be identified
- Opinion holder
  - All named entities, pronouns, and job titles are candidates for opinion holders.
- Opinion polarities
  - Opinion words
  - Relevance judgment

Evaluation

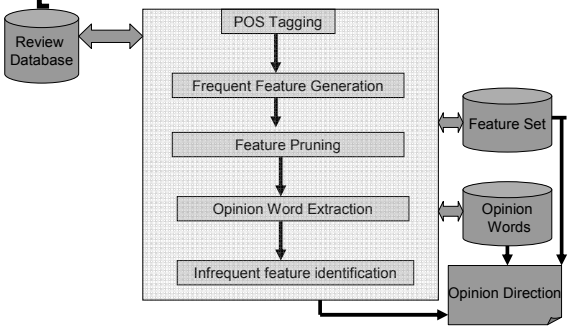
- Tested against human annotated corpus
- Precision = C/S
- Recall = C/T

  where
  - C = No of correct elements generated by system
  - S = Total number of elements generated by system
  - T = Total number of elements annotated by human

Challenges

- What are features mentioned?
- Opinion words such as **great, amazing, poor, bad** etc. indicate orientation of opinion
- Context sensitivity
  - **small** – may be positive or negative
  - **Too small** – negative
  - **Small but useful** – overall positive
  - **Reliable but expensive** - negative

Learning product features and opinion words



Extracting Product Features

- Pre-processing
  - Stemming
  - Stop word deletion
  - Fuzzy matching for word variant matches (auto-focus / autofocus/auto focus)
- Nouns and Noun Phrases from each sentence

Frequent Feature Generation

- Association rule mining for frequent item sets
- Features of a product are
  - *k*-item sets (k words occurring together)
  - Occur in at least *c%* of reviews

Compact Feature Sets

- Let *f* be a frequent feature
- Let *f* contain words $\{w_1, w_2, \ldots, w_n\}$ from sentence *s*
- *f* is compact in *s*, if $d(w_i, w_j) < \delta$ for all *i,j*
- If f occurs in m sentences and is compact in at least β of them – then f is compact

Redundancy pruning using *p-support* (Pure Support)

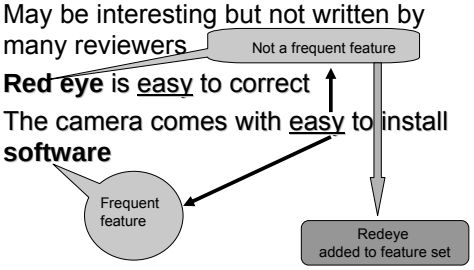
- Number of sentences that contain a feature as noun phrase and no feature phrase that is a superset of the feature
- Manual (10) , manual mode (4),manual setting (3) – different sentences
- *p-support* of manual = 3
- Frequent feature has defined *minimum p-support*

Extracting opinion words

- Extract adjectives near frequent features
- The strap is **horrible** and gets in the way of parts of the camera you need access to
- After nearly 800 pictures I have found that this camera takes **incredible** pictures

Infrequent features

- May be interesting but not written by many reviewers
- **Red eye** is easy to correct
- The camera comes with easy to install **software**



Context dependent opinion

- Intra-sentence conjunction
  - *Battery life is very long* – positive or negative??
  - The camera takes **great** pictures **and** has a **long** battery life - long is positive for battery
- **And** – positive conjugation in a sentence
- **But –** negative conjugation in a sentence
  - The camera takes **great** pictures **but** has a **short** battery life

Pseudo Intra-sentence conjunction

- Missing **and**
- Use opinion from other sentences like The camera has a long battery life which is **great**
- Contradictory reviews – majority counts

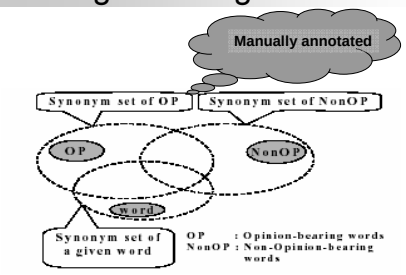
Inter-sentence conjunction

- The battery life is long and picture quality is great. However (But) the product is expensive.
- Great is positive so long is positive – but is negative – so expensive is negative

Learning opinion word orientation

- Seed list of opinion words from WordNet – manually annotate
- Synonyms and antonyms from WordNet
- Not all synonyms are opinion words
  - Good Ξ full, ample, solid

Learning to distinguish



Compute word's belongingness to Opinion and Non-Opinion Category

Membership computation

$\arg\max_c P(C|w) \approx \arg\max_c P(C|syn1, syn2, \ldots, synm)$

- where w is a target word
- C is category
- syn*j* is synonym or antonym of w

$$\arg\max_c P(C|w) = \arg\max_c P(C)\, P(w/C)$$
$$= \arg\max_c P(C)\, P(syn1, syn2, \ldots synm \mid C)$$
$$= \arg\max_c P(C) \prod_{k=1..m} P(f_k|C)^{count(f_k, synset(w))}$$

f_k = *kth* feature of category C which is also a member of the synonym set of the target word *w*
count(f_k, synset(w)) = total number of occurrences of f_k in the synonym set of *w*.

Alternately [Popescu & Etzioni]

- To identify semantic orientation of word *w*
- $P(l(w)=L|A_k)_{(m)} = P(l(w)=L|\bigcup_T(A_{k,T}))_{(m)}$
  - $A_{K,T}$ represents the labels assigned by A_k to neighbors of *w* connected to it by relationship (synonym/antonym) T
  - Ak = $\{w_j|L_j\}$ is the set of labeled word

Opinion Words

Table 1. Examples of opinion-bearing/non-opinion-bearing words

Adjectives	Final score	Verbs	Final score
Careless	0.63749	Harm	0.61715
wasteful	0.49999	Hate	0.53847
Unpleasant	0.15263	Yearn	0.50000
Southern	-0.2746	Enter	-0.4870
Vertical	-0.4999	Crack	-0.4999
Scored	-0.5874	combine	-0.5852

Opinion Aggregate Function [Ding and Liu]

- Sentence *s* contains multiple product features ($f_1, f_2, \ldots, f_m$) and multiple opinion words ($w_1, w_2, \ldots, w_n$)
- Orientation of opinion for each feature f_i in sentence *s*

$$score(f_i, s) = \Sigma_{w \in S} (w.SO/d(w, f_j))$$

where *w.SO* = Semantic Orientation of word *w*
$d(w,f_i)$ = distance between feature word and opinion word

OASYS 2.0 ("Opinion Analysis System")

- Developed in University of Maryland in collaboration with the University of Naples, Italy
- Identifies the intensity of opinion expressed by news sources
- Continuously crawls and analyzes news articles
- Identifies the topics in those articles and assigns a score to the intensity of opinion on the detected topics

Snapshot (

Opinion analysis from corpora *(Li et al., 2007)*

- MPQA and NTCIR-6
- SVM based system
- Opinionated sentence recognition
  - Each sentence classified as opinionated or not
- Opinion holder extraction
  - Opinion holder – one word or multiple words in opinionated sentence

Features used

- Unigram of tokens with two attributes – lemma and POS
- Sentence represented in terms of unigram - *tf* $*$ *idf*
  - *tf* - frequency of term in a sentence
  - *idf* - inverse of term frequency in all sentences in training data
- SVM algorithm for learning a two-class problem

Opinion holder extraction

- Features used (ANNIE)
- Every token's
  - Lemma
  - POS tag
  - Morphological features of word
  - Entity types

Syntactic Features

- Syntactic role using dependent tree output
  - Parser Minipar (http://www.cs.ualberta.ca/lindek/minipar.htm)
  - Stanford Parser
  - MontiLingua
  - Link Parser

Determining syntactic role

- Syntactic role of t where w is a child word;
  - whether w's parent chunk has an opinion word;
  - whether the chunk in which the word t is the head word in an argument of its parent chunk;
  - whether t has at least one child word and all its child words are in one (right or left) side of t in the sentence.

Trend analysis

- Blogosphere provides a a great source for trend analysis
  - product survey, customer relationship, and marketing
- Analyze the temporal trend in a group of blogs with common interests
- *Can reveal a lot of interesting trend information and insights*

Trend Analysis systems

- Collect information and identify patterns to predict the future
- Eigen Trend – identify trends from blogospheres
- Query trends analyzed demographically,chronologically ..

Google trends – temporal analysis of counts

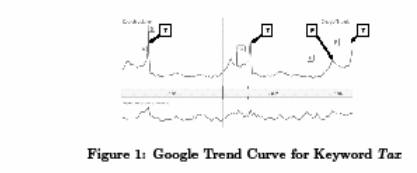


Figure 1: Google Trend Curve for Keyword *Tax*

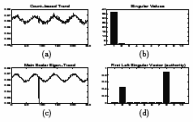
SVD based trend Analysis

- Given popularity vector of an item for n blogs over time window *j*
- $y_j = (y_{1j}, y_{2j}, \ldots, y_{nj})^T$
- $X=(y_1, y_2, \ldots, y_m)$ – for a keyword for m time periods
- Find trend vector $t = (t_1, t_2, \ldots, t_m)$

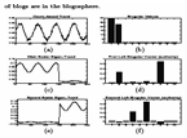
Scalar Eigen trend

- Solved using Singular Value Decomposition like HITS algorithm
- Assumption – a blog follows a certain distribution of keywords
- Distribution is determined iteratively

Sample performance



Count based trend analysis captures Overall trends but does not capture individual blog highs



Scalar Eigen Trend captures Separate interest in two blog groups

Conclusions

- Text Mining
  - key technology for analyzing and deriving intelligence from ever-increasing volumes of text-based information
- Research trends
  - Locate relevant information
  - Summarize
- Future trends
  - ***Foundation for next generation competitive and business intelligence platforms***
  - ***Predictive analysis – from actionable information to actionable intelligence***

Some relevant sites

- http://www.cs.pitt.edu/mpqa/
- http://oasys.umiacs.umd.edu/oasys/
- Y.Chi, B. Tseng, J. Tatemura, Eigen-Trend: Trend Analysis in the Blogosphere based on Singular Value Decompositions, CIKM 06.
- M. Hu and Bing Liu. Mining and Summarizing Customer Reviews. *SIGKDD 2004*.
- Soo-Min Kim and Eduard Hovy, Determining the Sentiment of Opinions. *Coling*, 2004.
- L.-W. Ku, L.-Y. Li, T.-H. Wu and H.-H. Chen. Major topic detection and its application to opinion summarization. *SIGIR 2005.*
- L-W Ku, Yong-Sheng Lo and Hsin-Hsi Chen, Using Polarity Scores of Words for Sentence-level Opinion Extraction, NTCIR-6, 2007.
- A-M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proc.HLT-EMNLP-05.*
- A. Esuli and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proc. CIKM-05.*
- E. Rillof, S. Patwardhan, J. Wiebe, Feature Subsumption for Opinion Analysis, EMNLP06.