Web Mining

Michalis Vazirgiannis
LIX @ Ecole Polytechnique

----

http://www.lix.polytechnique.fr/~mvazirg/

June 2014

Topics

- **Web personalization and recommendations (collaborative filtering)**
- Opinion Mining
- Web Advertising

Web personalization and recommendations

- ~25% of Internet users reading online reviews prior to paying for an offline service,
  - 80% claimed reviews had significant influence on their purchasing habits.
- Users pay a mark-up of 20% to 100% for services/products with excellent peer ratings on review sites.
- Humans are notoriously bad at choosing between too many choices,
  - rely on external recommendations and reviews to narrow the set of possible choices.

Personalization

- Personalized reviews tend to dominate
- Netflix: personalized video-recommendation system based on ratings and reviews by its customers.
- In 2006, offered a $1,000,000 prize to the first developer of a video-recommendation algorithm that could beat its existing algorithm

Recommender Data Model

- Set $U=\{u_1, \ldots, u_n\}$ of users
- Set $I=\{i_1, \ldots, i_m\}$ of items (e.g. products)
- Elements from U and I can be described by a vector respectively
  - $(a_1, \ldots, a_s)$ → attributes of user profile
  - $(b_1, \ldots, b_t)$ → description of items (meta data, features, …)
- Goal of recommendation process: recommend new items for an active user u
- Overview of process
  - User modeling (explicit or implicit, e.g. user rates items)
  - Personalization, generate list of recommended items

User-Item Ranking

- Recommendation often based on ratings of an item ij by a user u_k:
- Rating $r_{j,k}$: I → [0,1] ∪ ø
- Other range of values possible, e.g. {*, **, ***, ****, *****}
- ø := no rating for Item (or “0”)
- Example user-item matrix of ratings

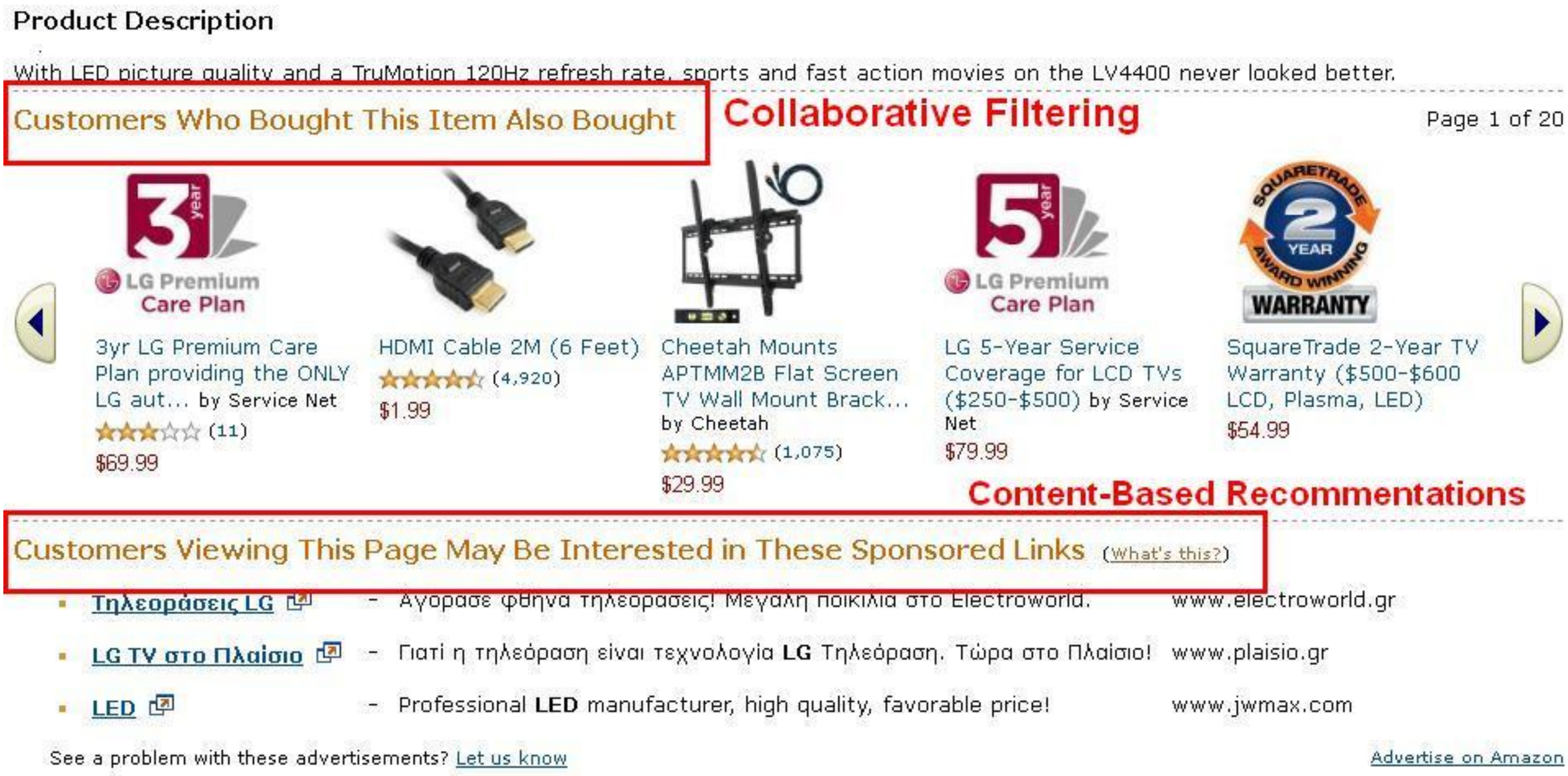
	V for Vendetta	La Vita e Bella	Lion King	Wall-e
Maria	4	3	2	4
Ted	ø	4	5	5
Andy	2	2	4	ø
Mike	3	ø	5	2

Types of Recommender Systems

- Collaborative filtering (CF)
- Content-based filtering (CB)
  - Individual recommender algorithms
  - Also utility- or knowledge-based approaches

- Case-based recommendation
- Hybrid recommender systems
  - Combination of several other recommenders
- Additional important variants
  - Context-aware and multi-dimensional recommenders
  - Decentralized recommender systems
  - Recommending for groups

Example: Product Page on Amazon



Issues of Recommender Systems

- Cold start and latency problems
- Sparseness of user-item matrix
- Diversity of recommendations
- Scalability
- Privacy and trust
- Robustness
- Utilization of domain knowledge
- Changing user interests (dynamics)
- Evaluation of recommender systems

Cold Start Problems

- “New user” and “new item” problem
- Systems cannot recommend items to new users with no profile or no interaction history
- Same for new items
  - Also “latency problem”: items need some time until they can be recommended
- Chicken-and-egg problem
  - Users will not use system without good recommendations
  - No incentive to rate items etc.
  - System cannot generate good recommendations
- Possible solutions
  - include explicit user profiling methods to start interaction

Data Sparseness

- Common situation
  - Lots of users and items
  - But only few ratings
  - Sparseness of user-item matrix
  - Recommender algorithms will not work very well
- In addition, new items are continuously added
  - Users should also rate these items
  - Number of ratings has to keep up with new users and items
- Possible solution
  - Include the automatic generation of ratings
  - Implicit user profiling, use of transaction history of users, e.g. click on a video constitutes a positive rating

Diversity of Recommendations

- Focus usually on generating recommendations as "good" as possible
  - But also important: new, unexpected items
  - Do not recommend items that are already known
  - Do not recommend items that are too similar to already known items
    - E.g. user likes "Lord of the Rings 1" → user possibly also likes "Lord of the Rings 2", but is this really a useful recommendation?
- Possible solutions
  - Use content-based approaches to easier integrate new items in recommendation process
  - Use collaborative filtering to allow "cross-domain" recommendations

Scalability

- Algorithms are based on matching users and items
  - The more items and users, the higher the computational effort to analyze the data
    - Storage/memory and runtime complexity
    - Alternatively, the quality of recommendations suffer
  - Scalability of recommender systems is an issue in practice
- Problem in particular with memory-based approaches
- Possible solutions include
  - Use model-based approach
  - Limit the number of items and/or users
    - E.g. only consider items that received at least k ratings
  - Pre-compute recommendations for users
    - Will reduce runtime

Privacy and Trust

- Collecting and interpreting personal data, e.g. ratings
  - For example, bought items or visited product Web pages on Amazon
  - Control for users?
    - Bought product may have been gift for other person
  - Privacy problem!
- Tradeoff with recommender quality
  - The more information about the user the system is able to collect, the higher the recommendation quality is in general
- Also trust, how can user trust the quality of a recommended item?
- Possible solutions include
  - Consider social relationships ("social recommender", "Web of Trust")
  - Let user control their profile information
  - Explanations of recommendations
    - Why was an item recommended?

Robustness

- Quality of (collaborative) recommenders depends on quality of ratings
  - Manipulation by users possible
    - E.g. by automatic registration of a large number of “users” and ratings
  - Also called “shilling”, “profile injection”
  - Attacks in principle
    - “push”: Aim is to push item(s) by inserting a large number of good ratings
    - “nuke”: Same with negative ratings
- Possible solutions include
  - Make registration for service harder, e.g. request and check personal information
  - Detect attacks and remove corresponding users and ratings
  - Adjust algorithms, some algorithms have proven to be more robust

Utilization of Domain Knowledge

- Systems often regard items in isolation
  - No relationships between items
  - No domain knowledge
- Example: searching for (books or other products on) “baseball”
  - Too many hits → restriction to “baseball technique”, or “baseball player”, for example
    - Based on user model and domain ontology
  - Too few hits → broading to “sport”, for example
- Some approaches in current research literature utilize Semantic Web technologies
  - Build and maintain item ontologies
  - Also for users
    - E.g. „GUMO“ (General User Modeling Ontology)

Changing User Interests (Dynamics)

- User model is often relatively static
- But dynamic evolution over user interests
  - Changes over time, older ratings may not be valid any more
- Also the context of recommendations
  - Example: Mobile restaurant guide
    - Restaurant may be too far away from current position (location)
    - Restaurant may be closed today (time)
  - A good rating for a restaurant after a dinner on a weekend may not be relevant for recommending a restaurant for a quick lunch on a workday
- Solutions in research literature include
  - E.g. explicit distinction between short- and long-term interests
  - Context-aware recommender systems

Evaluation of Recommender Systems

- Goal of personalization is to improve the interaction of users with the system
  - May be subjective, hard to evaluate
- General method for recommender systems
  - Let users rate recommended items and compare actual user ratings with predicted rating
  - Most important metrics
    - “precision”: probability rate that users did like recommended items
    - “recall”: probability rate that preferred items by users are recommended
  - In addition user studies
    - User evaluate system in questionnaire etc.

Recommender Systems

COLLABORATIVE FILTERING

Collaborative Filtering (CF)

- Basic idea: System recommends items which were preferred by similar users in the past
  - Based on ratings
    - Expressed preferences of the active user
    - And also other users → Collaborative approach
  - Works on user-item matrix
    - Memory-based or model-based
    - No item meta data etc.!
- Assumption: *Similar taste in the past implies similar taste in future*
- CF is formalization of “word of mouth“ among buddies

General Process

1. Users rate items
2. Find set S of users which have rated similar to the active user u in the past (→ neighborhood)
   - Similarity calculation
   - Select the k nearest users to the active user
3. Generate candidate items for recommendation
   - Items which were rated in neighborhood of u,
   - but were not rated by u yet
4. Predict rating of u for candidate items
   - Select and display n best items

Example (I)

	Hoop Dreams	Star Wars	Pretty Woman	Titanic	Blimp	Rocky XV
Joe	D	A	B	D	?	?
John	A	F	D		F	
Susan	A	A	A	A	A	A
Pat	D	A		C		
Jean	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Source: http://www.dfki.de/~jameson/ijcai03-tutorial/

Example (II)

	Hoop Dreams	Star Wars	Pretty Woman	Titanic	Blimp	Rocky XV
Joe	D	A	B	D	?	?
John	A	F	D		F	
Susan	A	A	A	A	A	A
Pat	D	A		C		
Jean	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Example (III)

	Hoop Dreams	Star Wars	Pretty Woman	Titanic	Blimp	Rocky XV
Joe	D	A	B	D	?	?
John	A	F	D		F	
Susan	A	A	A	A	A	A
Pat	D	A		C		
Jean	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Required Metrics

- Metric for user-user similarity
  - Mean-squared difference
  - Cosine
  - Pearson/Spearman correlation
- Select set S of most similar users (to active user u)
  - Similarity threshold
  - Aggregate neighborhood
  - Center-based
- Metric to predict the rating of u for an item i

Required Metrics

- Metric for user-user similarity
  - Mean-squared difference
  - Cosine similarity
  - Pearson/Spearman correlation
- Select set S of most similar users (to active user u)
  - Similarity threshold
  - Aggregate neighborhood
  - Center-based
- Metric to predict the rating of u for an item i

User-User Similarity

- Item set I
- Users U,V with u[i] denoting rating of item i by user u
  - the rating vector of user u is denoted by $\vec{u}$
  - the vector norm is denoted by $|\overrightarrow{u}|$
  - n is the number of items rated by both U and V

- Mean squared difference:
  - Small values show similar users $sim_1(U,V) = \frac{(\vec{u} - \vec{v})^2}{n}$

- Cosine similarity:
  - Large values show similar users $sim_2(U,V) = \frac{\overrightarrow{u} * \overrightarrow{v}}{|\overrightarrow{u}| * |\overrightarrow{v}|}$

Pearson/Spearman Correlation

- Average rating is taken into account
  - The vector of average ratings is denoted by $\overline{\overrightarrow{u}}$
- Not suitable for unary ratings
  - Unary: Item is marked (or not)
    - e.g. “Product was purchased“
  - Binary: good/bad, +/- etc.
  - Scalar: Numerical rating (e.g. 1-5) etc.
  - Consider only items which were rated by both users
- Values near 1 show similar users

$$sim_3(U,V) = \frac{(\overrightarrow{u} - \overline{\overrightarrow{u}}) * (\overrightarrow{v} - \overline{\overrightarrow{v}})}{|(\overrightarrow{u} - \overline{\overrightarrow{u}})| * |(\overrightarrow{v} - \overline{\overrightarrow{v}})|}$$

Example Calculation

User/item	a	b	c	d	e	f
U	5		3		4	
A	1	1		1		
B	1		3	1		
C	5	2	2		5	4
D		3		2		

$Sim_1(U,V)$	$Sim_2(U,V)$	$Sim_3(U,V)$
-	-	-
16	1	0
8	0.76	-1
2/3	0.98	0.833
∞	∞	∞

Required Metrics

- Metric for user-user similarity
  - Mean-squared difference
  - Cosine
  - Pearson/Spearman correlation
- Select set S of most similar users (to active user u)
  - Similarity threshold
  - Aggregate neighborhood
  - Center-based
- Metric to predict the rating of u for an item i

Neighborhood of Similar Users

- Goal: Determine set S of users which are most similar to the active user u
- Center-based
  - S contains k most similar users
    - Problem: maybe some of the users are not really that similar, if k was chosen too large, deviators possible
- Similarity threshold
  - S contains all users with a similarity bigger than a threshold t
    - Problem: maybe too few users in S
- Aggregate neighborhood
  - Follow similarity threshold method first
  - If S is too small (less than k users)
    - Determine "centroid" of set S and add users which are most similar to centroid (→ less deviators than center-based method)

Required Metrics

- Metric for user-user similarity
  - Mean-squared difference
  - Cosine
  - Pearson/Spearman correlation
- Select set S of most similar users (to active user u)
  - Similarity threshold
  - Aggregate neighborhood
  - Center-based
- Metric to predict the rating of u for an item i

CF Recommender (I)

- Given
  - Set S with most similar users to u
  - s[i] rating of a user (from S) from an item i
- Goal: Predict the rating of u for i
- Easiest option: Arithmetic mean

$$r_1(U, i) = \frac{1}{|S|} \sum_{s \in S} s[i]$$

- Problems
  - Similarity of u with members of S is not taken into account
    - Solution: Weighting based on similarity

CF Recommender (II)

- Different users utilize rating scale differently
  - Solution: Consider deviation from average rating (for user)

$$r_3(U,i) = \overline{u} + \frac{1}{\sum_{s \in S} sim(U,s)} \sum_{s \in S} sim(U,s) * (s[i] - \overline{s})$$

- Note
  - Many variations of algorithms in research literature
    - For various application domains, with different properties

Collaborative Filtering

- Amazon and other commercial service use some form of collaborative filtering
  - Exact method usually not published
- Non-commercial example with published algorithms: http://www.movielens.umn.edu
- Exercise ☺
  - Comprehend calculation for introductory example
  - Substitute 1:=A, 2:=B etc.
  - Calculate predicted rating of user “Joe“ for movies “Blimp“ and “Rocky XV”

Advantages Collaborative Filtering

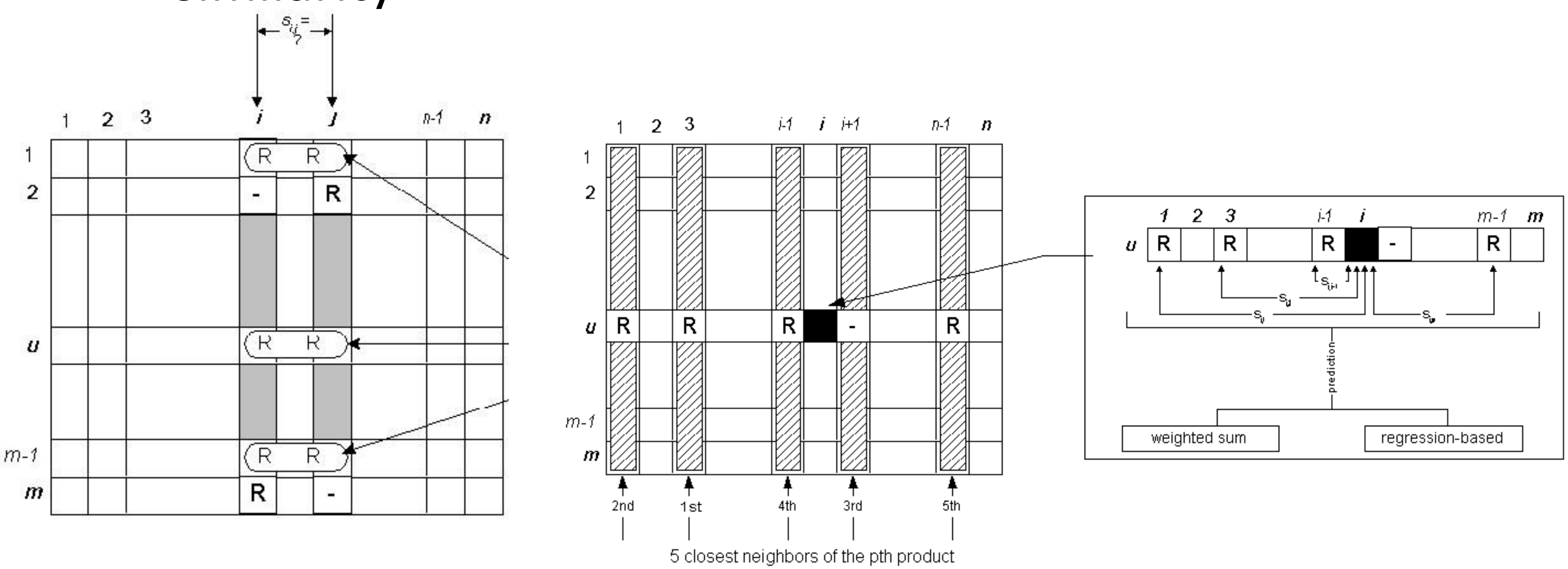
- Works well in practice
- Quality of recommendations improves with density of ratings
- Only ratings as input data required
  - In particular, no information (meta data, description) about items needed
- CF is able to generate cross-domain ("cross genre") recommendations → high diversity
  - Because item categories etc. are not considered
  - Has proven useful in practice
- Implicit user feedback often adequate (CTR)
  - Unary ratings, e.g. rating = "Click on product Web page"

Disadvantages Collaborative Filtering

- New user and new item problem
  - Serious issue in practice
- Often sparseness in user-item matrix
  - Algorithms generate worse results with too few ratings
- “Grey sheep” problem
  - Does not work very well for users with “extraordinary” taste
    - Because similar users are not available
  - Also “black sheep”, users that intentionally make incorrect ratings
    - CF is prone to manipulation
    - Trust and robustness are issues

Item-to-Item Collaborative filtering (Amazon)

- Item representation through a N-dimensional vector.
  - Each dimension corresponds to a user's action on this item.
- Rather than matching the user to similar customers, build a *similar-items table by finding that customers tend to purchase together.*
- Recommend items with high-ranking based on similarity



References

- presentation inspired by slides of Wolfgang Wörndl for "User Modeling, Personalization and Recommender Systems" course in Technical University Munich

  http://www11.in.tum.de/Veranstaltungen/CSCW2:UserModeling,PersonalizationandRecommenderSystems%28IN2119%29
- D. Billsus and M. J. Pazzani, "Learning collaborative information filters", In Proceedings of the Fifteenth International Conference on Machine Learning, pages 46{54, 1998
- "A Comparative Study of Collaborative Filtering Algorithms", Joonseok Lee, Mingxuan Sun, Guy Lebanon, http://arxiv.org/pdf/1205.3193.pdf
- A. Paterek. Improving regularized singular value decomposition for collaborative filtering, Statistics, 2007:2{5, 2007}.

Topics

- Web personalization and recommendations (collaborative filtering)
- **Opinion Mining**
- Web Advertising

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Motivation

- “What other people think” has always been an important piece of information for the decision-making process:
  - Who to vote
  - Decide which dishwasher to buy
  - Decide the next holiday destination

- Internet & Web made it possible to find out about opinions and experiences of those in vast pool of people:
  - 81% of Internet users have done online research on a product at least once
  - 20% do so on a typical day

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Terminology

- "the ideal **Opinion Mining** tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)" [Dave et al. 2003]

- **Sentiment Analysis** synonym to **Opinion Mining:**
  - "Sentiment" used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments [Das and Chen]
  - The Sentiment Analysis term is construed more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Applications

① **Review-related websites**: creation and automated upkeep of review and opinion aggregation websites

② **As a sub-component technology:** enabling technologies for other systems (e.g. recommendation systems → not recommend items that receive negative feedback)

Applications

③ **Business and government intelligence:**

- Reasoning for high or low sales
- Monitoring sources for increases in hostile or negative communications

④ **Application across different domains:**

- **Politics:** understand what voters are thinking [Mullen and Malouf 2006] / classification of politicians' positions [Thomas, Pang and Lee 2006]
  - **Sociology:** who is positively or negatively disposed towards whom → who would be more or less receptive to new information transmission from a given source [Rogers]

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Challenges

- Contrast with fact-based textual analysis

    - text categorization classifies documents by topic

    VS

    sentiment classification involves few classes (e.g. “positive” or “3 stars”)

    - different characteristics of answers to opinion-oriented questions compared to fact-based questions

Challenges

- Factors that make opinion mining difficult:

	Proposed word lists	Accuracy
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58%
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64%
Statistics-based	positive: *love, wonderful, best, great, superb, still, beautiful* negative: *bad, worst, stupid, waste, boring, ?, !*	69%

l.

Challenges

- Non-trivial to recognize opinion holders

Consider the following quote from Charlotte Brontë, in a letter to George Lewes:

You say I must familiarise my mind with the fact that "Miss Austen is not a poetess, has no 'sentiment' " (you scornfully enclose the word in inverted commas), "has no eloquence, none of the ravishing enthusiasm of poetry"; and then you add...

→ the opinion is not that of the author, but the opinion of "**You**", which refers to **George Lewes** in this particular letter

Challenges

- Order effects can completely overwhelm frequency effects

*This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a **good** performance. <u>However, it can't hold up</u>.*

→words that are positive in orientation dominate this excerpt and yet the overall sentiment is negative because of the crucial last sentence

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Classification

- **Sentiment polarity classification:**
  - classify the opinion as falling under *one of two opposing sentiment polarities*, or
  - locate its position on the continuum between these two polarities
  - Examples:
    - reviews: thumbs up/ thumbs down
    - prediction in elections: likely to win/ unlikely to win

- **Subjectivity detection and opinion identification**

  decide whether a given document contains subjective information or not (facts), or identify which portions of the document are subjective

  - Yu et. al. achieve high accuracy (97%) with a Naive Bayes classifier on a particular (Wall Street Journal articles),
  - task is to distinguish News and Business (facts) articles from Editorial and Letter to the Editor (opinions) articles

Classification

- Joint topic-sentiment analysis

  interactions between topic and opinion that make it desirable to consider the two simultaneously

- Viewpoints and perspectives

  Much work on analyzing sentiment and opinions in politically-oriented text focuses on general attitudes expressed through texts that are not necessarily targeted at a particular issue or narrow subject

Opinion Mining

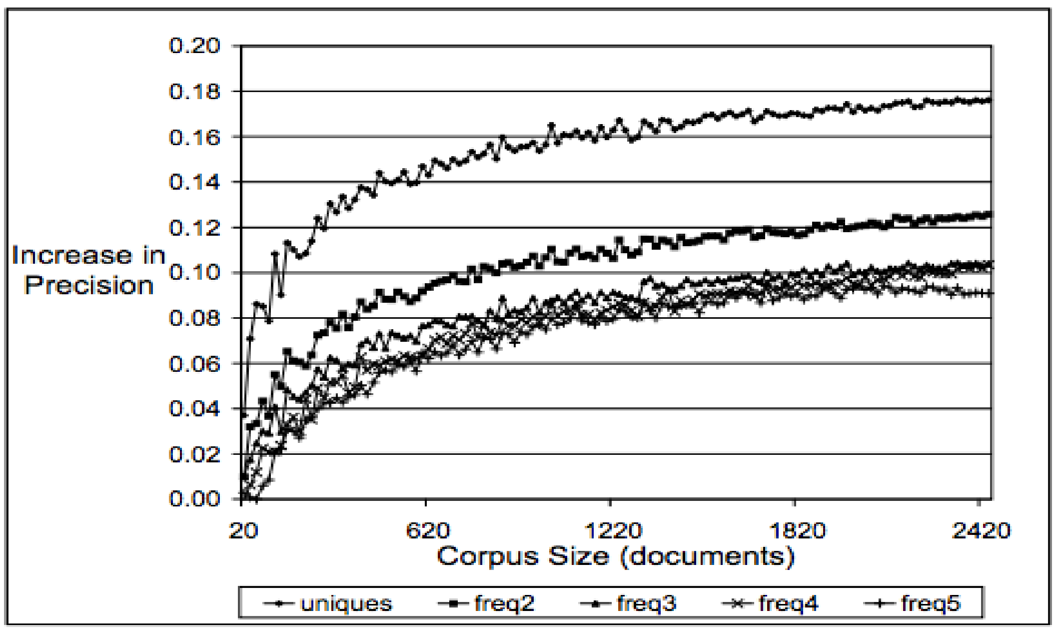
- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Feature selection

→Converting a piece of text into a feature vector

①Term presence vs. frequency

Pang et al. obtained better performance using *presence* rather than *frequency* (entries of vector indicate whether a term occurs (value 1) or not (value 0) VS tf-idf weighting)



[Wiebe, Wilson, …2004]

Feature selection

② Term-based features beyond term unigrams

The *position* of a token within a textual unit (e.g., in the middle vs. near the end of a document) can have effects on how much that token affects the overall sentiment or subjectivity status of the enclosing textual unit

Table 4: Reason sentence identification results on restaurant reviews.

Features used	Acc (%)	Prec (%)	Recl (%)	F-score (%)
Op	61.64	60.76	47.48	53.31
Lex	63.77	67.10	51.20	58.08
Lex+Pos	**63.89**	67.62	51.70	58.60
Lex+Op	61.66	**69.13**	**54.30**	**60.83**
Lex+Pos+Op	63.13	66.80	50.41	57.46
Baseline	**54.82**			

Table 2: Feature summary.

Feature category	Description	Symbol
Lexical Features	unigrams bigrams trigrams	***Lex***
Positional Features	the first, the second, the last, the second to last sentence in a paragraph	***Pos***
Opinion-bearing word features	pre-selected opinion-bearing words	***Op***

[Kim and Hovy 2006]

Feature selection

③ Parts of speech (POS)

- Work on subjectivity detection revealed a high correlation between the presence of adjectives and sentence subjectivity [Hatzivassiloglou and Wiebe 2000]

- the baseline of using just attitude-bearing adjectives is reasonably high [Whitelaw, Garg and Argamon 2005]

④ Syntax

→ incorporating syntactic relations within feature sets

- a subtree-based boosting algorithm using dependency-tree-based features outperformed the bag-of-words baseline [Kudo and Matsumoto 2004]

Feature selection

⑤Negation

While the bag-of-words representations of “I like this book” and “I don’t like this book” are considered to be very similar by most commonly-used similarity measures, the only differing token, the negation term, forces the two sentences

ID	Approach	Selected Terms	Term Weighting	DF	Terms labeled with POS tags	Negation	Accuracy
3	Unigram with TFIDF	All	TFIDF	3	No	No	76.50%
4	Unigram with TFIDF and DF = 1	All	TFIDF	1	No	No	74.17%
5	Unigram labeled with POS	All	TFIDF	3	Yes	No	75.83%
1	Unigram with negation phrase and DF = 3	All	TFIDF	3	No	Yes	78.33%
2	Unigram with negation phrase and DF = 1	All	TFIDF	1	No	Yes	79.33%

[Na, Sui, Khoo, Chan and Zhou 2004]

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Approaches

① Supervised learning

- **classification**

Naïve Bayes, Maximum Entropy and Support Vector Machines achieve accuracies form ~77% to ~82% while using different combinations of features (performance not as good as in topic –based categorization) [Pang et al.]

- **regression**

research that explicitly considers regression or ordinal-regression formulations of opinion mining problems → "how positive is this text?" and "how strongly held is this opinion?" [Goldberg and Zhu 2006]

Approaches

② Unsupervised approaches

- **Unsupervised lexicon induction:** a sentiment lexicon is firstly created in an unsupervised manner, and then the degree of positivity (or subjectivity) of a text unit is determined via some function based on the positive and negative (or simply subjective) indicators, as determined by the lexicon, within it

- **Bootstrapping:** use the output of an available initial classifier to create labeled data, to which a supervised learning algorithm may be applied [Riloff and Wiebe 2003]

Approaches

③Classification based on relationship information

- treating a document as a bag of features, then model the structure of a document via analysis of sub-document units, and explicitly utilize the relationships between these units, in order to achieve a more accurate global labeling

- texts from a running discussion represents a rich information source that references between such texts that can be exploited for better collective labeling of the set of documents

Approaches

④ Incorporating discourse structure

discourse structure (e.g. twists and turns in documents) tends to have more effect on overall sentiment labels [Pang et al.]

④ Language models

mechanisms for assigning probabilities to text rather than labels drawn from a finite set (cannot be defined as either supervised or unsupervised classifiers) [Eguchi and Lavrenko 2006]

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Summarization

→aggregating and representing sentiment information drawn from an individual document or from a collection of documents

① Single-document opinion-oriented summarization

- approaches that create textual sentiment summaries based on extraction of sentences or similar text units
- other methods can work directly off the output of opinion-oriented information-extraction systems

① Multi-document opinion-oriented summarization

- challenge: determining which documents or portions of documents express the same opinion

Opinion Mining

- Motivation
- Terminology
- Applications
- Challenges
- Technology
  - Classification
    - Feature selection
    - Approaches
  - Summarization
- Broader Implications

Broader implications

- **Privacy**

  applications that gather data about people's preferences can trigger concerns about privacy violations

- **Manipulation**

  corporations planting positive reviews or attempting to use untoward means to manufacture an artificially inflated reputation or suppress negative information

- **Economic impact of reviews**

  how much effort companies might or should want to expend on online reputation monitoring and management?

References

- Pang, B., &Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.
- Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003
- Das, S., & Chen, M. (2001, July). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific finance association annual conference (APFA) (Vol. 35, p. 43).
- Mullen, T., & Malouf, R. (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 159-162).
- Thomas, M., Pang, B., & Lee, L. (2006, July). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 327-335). Association for Computational Linguistics.
- Rogers, E. M. (2010). Diffusion of innovations. Simon and Schuster.
- Kudo, T., & Matsumoto, Y. (2004, June). A Boosting Algorithm for Classification of Semi-Structured Text. In EMNLP (Vol. 4, pp. 301-308).

References

- Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.
- Hatzivassiloglou, V., & Wiebe, J. M. (2000, July). Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the 18th conference on Computational linguistics-Volume 1 (pp. 299-305). Association for Computational Linguistics.
- Goldberg, A. B., & Zhu, X. (2006, June). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing (pp. 45-52). Association for Computational Linguistics.
- Riloff, E., & Wiebe, J. (2003, July). Learning extraction patterns for subjective expressions. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 105-112). Association for Computational Linguistics.
- Whitelaw, C., Garg, N., & Argamon, S. (2005, October). Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 625-631). ACM.