DYNAMIC SUMMARIZATION OF MICROBLOG STREAMS

Younos Aboulnaga
University of Waterloo

13 November, 2012

UNIVERSITY OF Waterloo

(or any Time Series of Short Non-canonical Text)

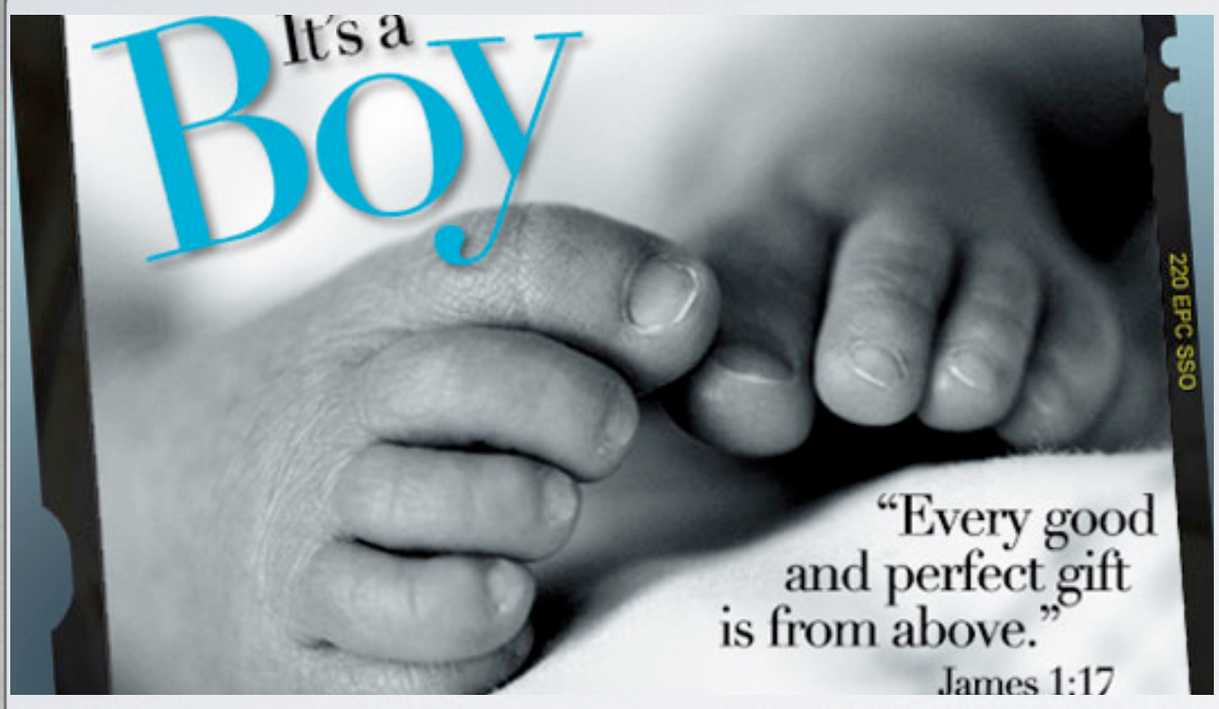
CURRENT SOCIAL MEDIA

Like being in a crowded bar and a protest at the same time

(you might also have many news agencies as "friends")

UNIVERSITY OF Waterloo

NOT PERSONAL ANYMORE



world. The Kayapo being expelled from their homes for the construction of the Belo Monte Dam, which will flood 400.000 acres of the Amazon Rainforest in Brazil."

Nauseating juxtaposition of social content and world news!

Features of Social Media emphasizes on the Media aspect

The Social aspect is left for the user to maintain

UNIVERSITY OF Waterloo

VISION

- Give users the best of the two worlds, Social Networking and Social Media, by separating them into two streams
  - A *Social stream* of content shared with friends only
    - Easy to maintain since "friends" are really friends
  - A *Media stream* of content shared publicly
    - Includes posts of news agencies, companies PR, .. etc
    - Makes use of the social network for selecting content

UNIVERSITY OF Waterloo

4

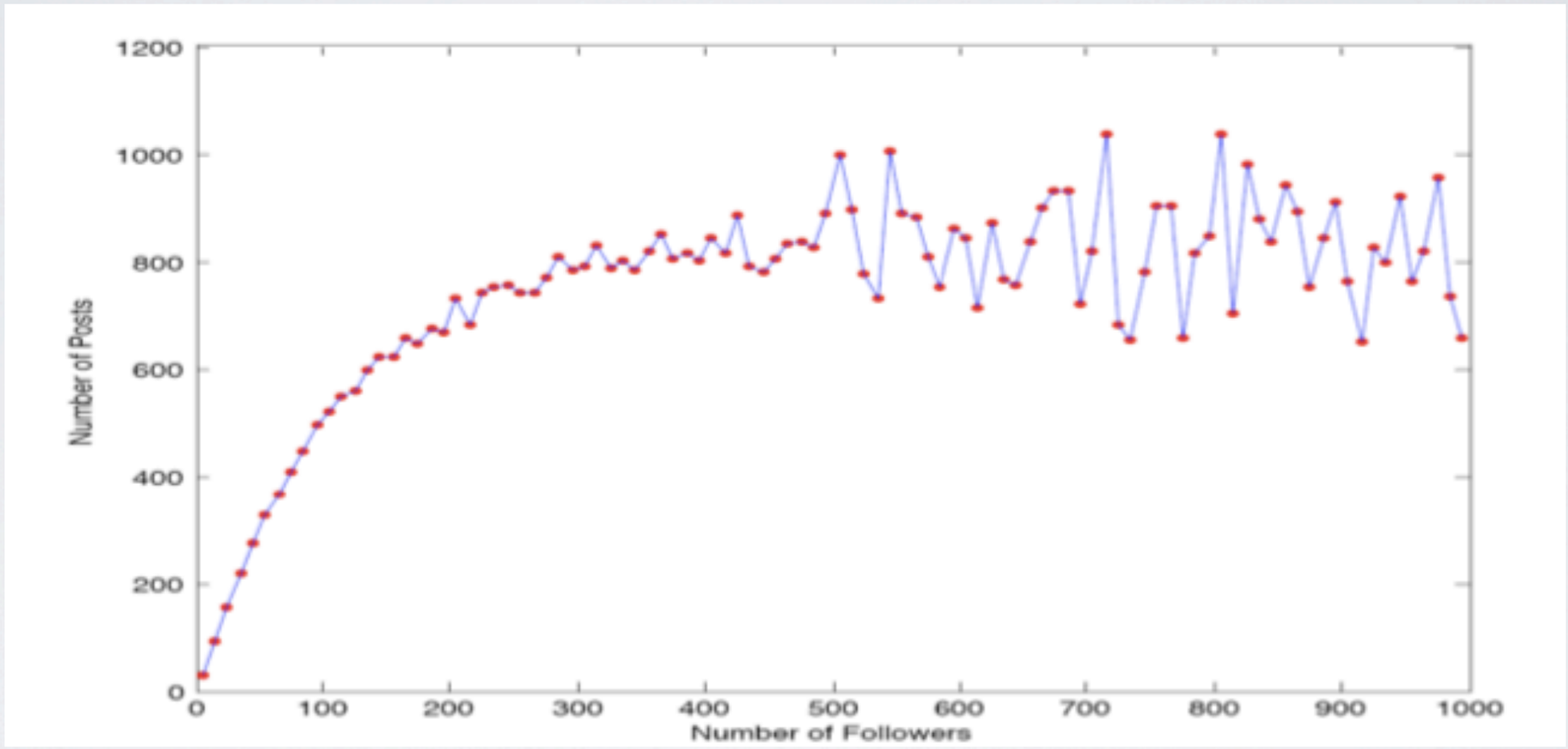
GOALS

- Automatic separation of Media and Social content
  - Using content features rather than source features
- Browsable timeline of the Media stream
  - Temporal summary of open domain topics
  - Capturing bursty/trending topics as well as those with long runs of slightly higher than normal interest
- Using simple techniques that scale well

UNIVERSITY OF Waterloo

sustained

THE COME BACK OF JOURNALISM AND PR

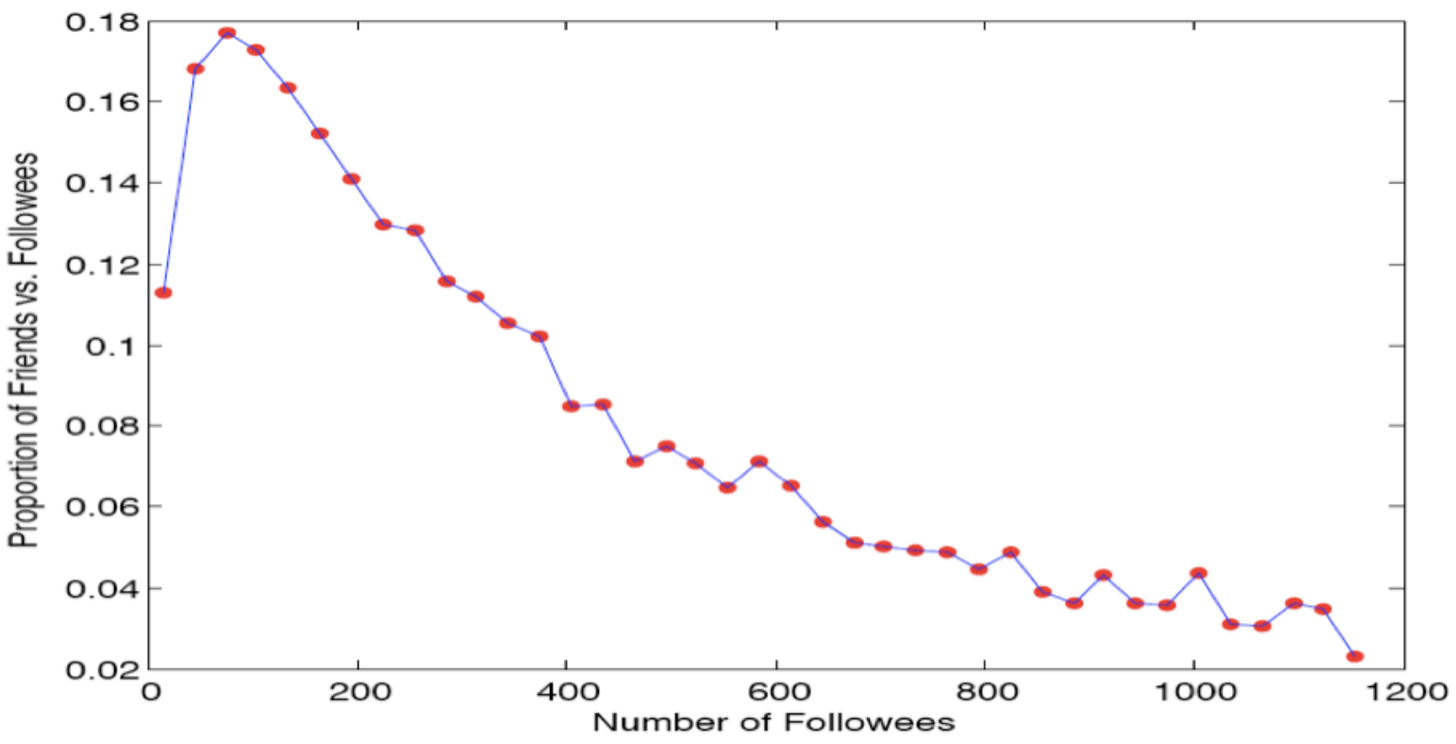


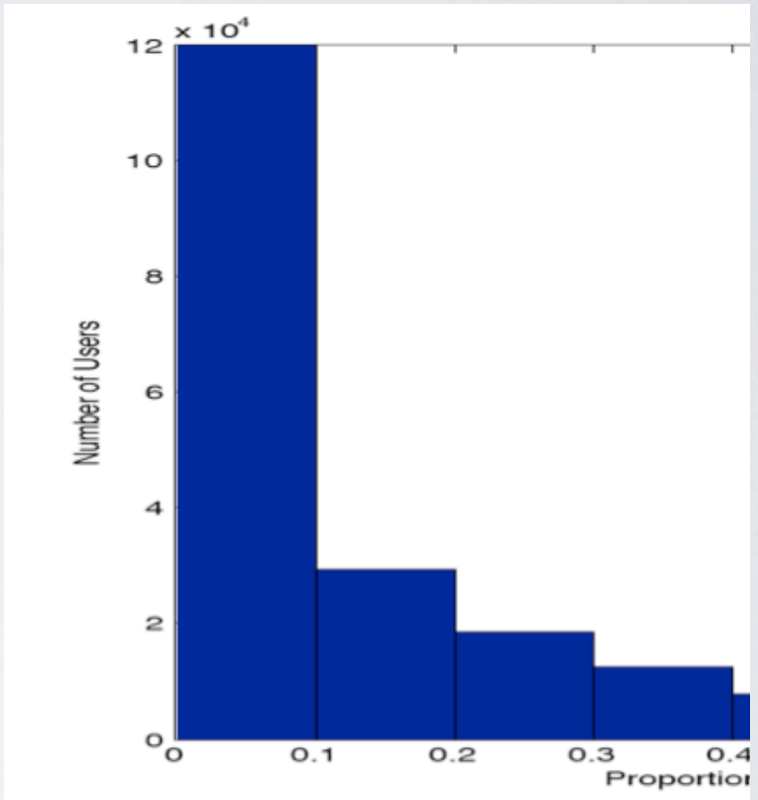
In a dataset of 309740 users collected by HP, the average user had 85 followers and posted 255 posts. (Huberman et al, 2009)

For each user of Twitter in our data set we obtained the number of followers and followees (people followed by a user) the user has declared, alongwiththecontentanddatestampofallhisposts.1 Ourdatasetconsisted of a total of 309,740 users, who on average posted 255 posts, had 85 followers, and followed 80 other users. Among the 309,740 users only 211,024 posted at least twice. We call them the active users. We also define the active time of an active user by the time that has elapsed between his first and last post. On average, active users were active for 206 days.

Around 25.4% of all posts are directed, which shows that this feature is widely used among Twitter users.

THE COME BACK OF JOURNALISM AND PR





Defining a user's friend as a person to whom the user has directed at least two posts, people aren't following friends!

Stress that the definition is somewhat loose one.

Around 25.4% of all posts are directed, which shows that this feature is widely used among Twitter users.

A TEXT MINING VIEW OF SOCIAL MEDIA

- A time series of documents
  - Not strictly a stream: documents are stored and deletion is allowed
- Documents are short (except in case of Blogs)
- Mixed language, code switching and romanization
- Language is non-canonical (also sarcastic in many cases)
  - Accuracy of syntactic parsing of text from Google Web Treebank is only in the 80-84% range. (Petrov and McDonald, 2012)

UNIVERSITY OF Waterloo

Explain:
NLP
Tree Bank (Wall Street Journal vs Web)
How is it related to my stuff

CHALLENGES

- High arrival rate of documents (thousands per second)
- Near real time update of index and/or models required
- Evolving vocabulary with a non-stationary distribution
  - In Twitter, comparing the *10000* highest frequency terms between two consecutive hours, an average of *1004* terms are replaced, and *15* of the new terms are out of vocabulary terms (never seen before). (Lin and Mishne, 2012)

UNIVERSITY OF Waterloo

OPPORTUNITIES

Imitation (Leskovec, 2009)

- Important content is shared by many sources/users
  - Literal collaborative filter of good content
- Different wording and various points of view
- Assumption: Topical words remain the same
  - Named Entity + piece of news or opinion

- Hashtags (caveat only 10% of Tweets are tagged)

IMITATION
An Example

- Topic "Pakistan diplomat arrest murder"
  - BBC News - US diplomat charged with Pakistan double murder
  - DTN India: US mounts pressure on Pakistan to release 'illegally detained' murder-accused diplomat: Islamabad, Fe... [link]
  - Case Of Jailed Diplomat In Pakistan Fuels Anger [link]
  - Official to face murder rap in Pakistan, fury against US grows
  - The Mystery Deepens: Was American Arrested In Pakistan 4 Killing 2 Men An Assassin?—US Claims He Has 'Diplomatic Immunity'

UNIVERSITY OF Waterloo

FREQUENT ITEMSET MINING

A.K.A. Frequent Pattern Mining

- Simple and well studied mining technique
- Suitable for short documents
- Parallel and distributed implementations available
- Highly adopted in various stream settings
- Introduction of new vocabulary terms is not a problem

UNIVERSITY OF Waterloo

12

Parallel and Distributed = Map/Reduce

FREQUENT ITEMSET MINING

With candidate generation

- Support is the number of times an itemset appears
  - A set of items cannot have higher support than any of its subsets (downward closure)
    - Therefore itemsets with support lower than a certain threshold need not be expanded any further
- Candidate generation is the bottom up generation of frequent itemsets combinations
  - The bottle neck of Apriori (Agrawal, 1994) like algorithms

graphical

FREQUENT ITEMSET MINING

Without candidate generation

- The candidate generation step can be skipped by representing the database as a tree structure

- FP-Growth (Han, 2000) constructs a prefix tree of frequent itemsets co-occurring with each item individually

  - The search for itemsets happens by a partitioning-based, divide-and-conquer method

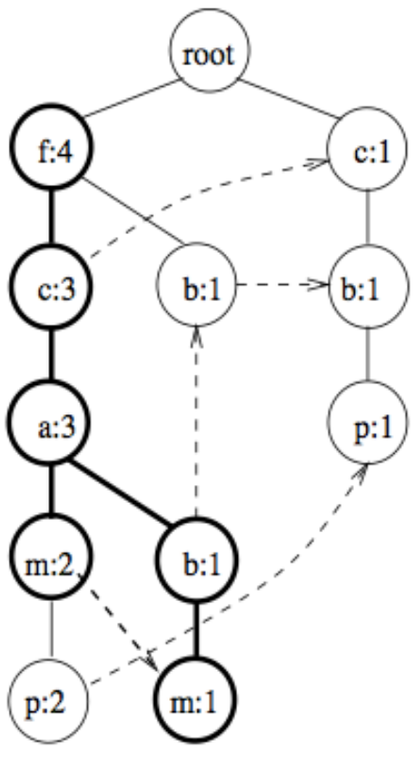
  - PFP-Growth (Li et al., 2008) is a Map/Reduce adaptation of FP-Growth that achieves near linear scalability

UNIVERSITY OF Waterloo

To quantify speedup, we took 100 ma- chines as the baseline and made the assumption that the speedup when using 100 machines is 100, compared to using one machine. This assumption is reasonable for our experi- ments, since our algorithm does enjoy linear speedup when the number of machines is up to 500. From Figure 4, we can see that up to 1500 machines, the speedup is very close to the ideal speedup of 1:1. As shown in the table attached with Figure 4, the accurate speedup can be computed as 1920/2500 = 76.8%. This level of scalability, to the best of our knowledge, is far better than previous attempts [10, 8]. (We did not use the same data set as that used in [10, 8] to perform a side-by-side comparison. Nevertheless, the substantial overhead of these algorithms hinder them from achieving a near-linear speedup.)

FP-GROWTH ILLUSTRATION

Conditional FP-Tree for "m"



Global FP-tree

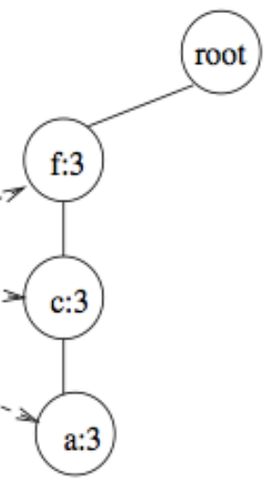
(f:2, c:2, a:2)

(f:1, c:1, a:1, b:1)

Conditional pattern base of "m"

Header table

item	head of node-links
f	
c	
a	



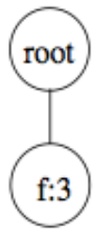
Conditional FP-tree of "m"

Conditional pattern base of "am": (f:3, c:3)

Conditional FP-tree of "am"

Conditional pattern base of "cm": (f:3)

Conditional FP-tree of "cm"



15

FIM IN MICROBLOG DATA

- Will FIM work on social media content?
  - Proof of concept: Using FIM to improve retrieval performance on the TREC 2011 Microblog track.

- Text Retrieval Evaluation Conference (TREC) is an annual information retrieval conference and competition
  - 59 groups from across the world participated in the 2011 Microblog track, with 184 submissions (Ounis, 2011)

UNIVERSITY OF Waterloo

TREC 2011 MICROBLOG DATASET

- About 16 millions tweets gathered in the period from January 24th to February 9th, 2011
- 49 queries with known relevant tweets
- Patterns in the corpus were mined using Mahout's implementation of PFP-Growth, items are unigrams
  - The top one percentile of unigrams, in terms of frequency, were discarded after the counting step

UNIVERSITY OF Waterloo

17

. Give example that "the" can and will appear with any other noun, and we don't want to mine and asses
. Why top 1?

PERFORMANCE ON TREC MICROBLOG 2011

- Precision at 30 (the track's measure of choice in this year) of our technique and our baseline are shown below, along with other landmark submissions.
  - With FIM query expansion we beat 2011's best, with a difference that is statistically significant at the 99% confidence level

2011 90th Percentile	Baseline BM25	2011 Best	FIM Query Expansion
0.4095	0.4177	0.4551	**0.4626**

UNIVERSITY OF Waterloo

18

FREQUENT ITEMSETS EXAMPLES

Same support - Same time period

- Informative: [#neversaynever, never, see, bieber, #nmlmademeabelieber, would, neversaynever, music, new, thought, live, tuesday, justin]
- Bots: [active, moon, house, leo, other, today, aquarius, 7th]
- Language constructs: [too, really, much]

UNIVERSITY OF Waterloo

19

Support: 162
Period:

QUERY EXPANSION

- Adding terms to the query
  - Increases diversity and specificity
  - Can suffer from concept drift
- Pseudo Relevance Feedback
  - Run the original query against the documents index
  - Create a language model of the top results
  - Choose expansion terms from the model

QUERY EXPANSION USING FREQUENT ITEMSETS

- Itemsets are indexed
- For each query:
  - The query is used to search the frequent itemsets
  - Terms from itemsets with the highest relevance score are added to the query, until *n* terms are added
  - Expansion terms are weighted uniformly

UNIVERSITY OF Waterloo

21

FIM QUERY EXPANSION EXAMPLE

- Frequent itemsets for topic "Pakistan diplomat arrest murder"
  - [american, city, diplomat, killed, kills, lahore, pakistan, pakistanis, shot, two] - score: 19.47
  - [consulate, employee, pakistan, armed, killed, shot] - score: 9.06
  - [pakistan, drone] - score: 9.06
  - [murder, sonawane] - score: 8.55
  - [#tunisia, arrest, tunisia, warrant] - score: 8.52

22

Say how stemming can cause concept drift; e.g. Egyptian -> Egypt

ADAPTING FIM TO NATURAL LANGUAGE

- FIM on Natural Language results in many itemsets that are non-informative language constructs
  - High support
  - Not necessarily be made up of stop words
  - Examples: [post, new, blog], [friend, best], [video, youtube, upload], [least, at], [NON_ENGLISH_TOO]

SUBJECTIVE INTERESTINGNESS

- Chooses itemsets based on the strength of association
- Studies the interrelation between itemsets
  - To derive rules not normally captured by association rule mining; for example, competition (Liang et al., 2009), and
  - To reduce redundancy and discover "interesting patterns". (Brin et. al, 1997 and Lee et al, 2003)
- We also study the interrelation along the time dimension

UNIVERSITY OF Waterloo

MEASURES OF ASSOCIATION

- Normalized Mutual Information
  - Calculated between the whole Itemset *I* and the head term *h* for which itemsets are mined

$$NMI(I,h) = \frac{\sum_{t \in I} p(t,h) \ln \frac{p(t,h)}{p(t)p(h)}}{-\sum_{t \in I} p(t,h) \ln p(t,h)}$$

  - Joint probability can be estimated from conditional FP-Tree
  - Used in the TREC runs

UNIVERSITY OF Waterloo

25

Stress that NMI is using Itemset support

MEASURES OF ASSOCIATION

- NMI doesn't take into account the probability of one term but not the other; what is the surprise in the pair?
- Measures of correlation for contingency table
- Calculated between pairs of items from the Itemset *I*, then averaged using the adequate mean
- Many measures are parametric (make assumptions)
- Working on finding a measure for the NxN case

Spearman's Correlation

Kendall's Tau

Goodman – Kruskal Gamma

Chi–square (c^2)

It should be noted that the c^2–test is quite sensitive to the sample size. If the sample size is too small, the c^2 value is overestimated; if it is too large, the c^2 value is underestimated. To overcome this problem, the following measures of association are suggested in the literature: *Phi–square* (j^2), *Cramer's V* and Contingency Coefficient.

Fisher's Exact Test

Yule Q

CONTINGENCY TABLE EXAMPLE

	Bieber=1	Bieber=0	Sum
Married=1	Low	Medium	
Married=0	High	High	
Sum			

- Unexpected associations are the most informative
  - Rarely occur together, but not negatively correlated
  - Easier to calculate positive/negative correlation

27

DYNAMIC FIM

- The dynamic nature of social media streams is useful for detecting interesting topics through change in correlation
- Online/Incremental and stream algorithms are an overkill since only near realtime updates are required
  - Block evolution (Ganti et al, 2002) is suitable as long as a block can be mined before the next block arrives
  - By Map/Reduce's scalability any block length is possible

UNIVERSITY OF Waterloo

28

BLOCK EVOLUTION

- Proposed by Ganti et al. (2002), this is a very practical approach for many mining applications where the real time constraint is relaxed

- Updates are received in blocks (batches) of documents that occurred within a fixed duration

- Blocks are non-overlapping, limiting the possible window lengths to multiples of the duration of one block

- Requires storing partial models from all blocks

- Addresses the need for deleting records.

DETECTING BURSTY TOPICS

- Bursty topics are mined from short blocks
- Itemsets for an hour in 24 Jan. 2011, using 15 min blocks

NMI	Correlation
[coast, east],[moscow, bomb], [moscow, airport],[35, russia], [speech, obama],[engine, search], [coffee, starbucks], [according, sources],[airport, blast], [state, union]	[moscow, airport], [#constantcontact, via], [coast, east],[ly, bit],[moscow, bomb],[estate, real],[state, union],[york, new], [channel, subscribe, youtube]

UNIVERSITY OF Waterloo

A bit about why 15 minutes

THE RIGHT BLOCK AND WINDOW LENGTH

- Very short blocks do not provide enough support to informative itemsets; only language constructs are mined
  - That is, if the partial model to store is beyond counts
- Longer blocks mean that a window advances more slowly
- Itemsets mined from a longer block are not the same as the concatenation of itemsets mined from shorter blocks
  - Is it possible to merge itemsets mined from blocks?

UNIVERSITY OF Waterloo

DETECTING NON-BURSTY TOPICS

- Topics with sustained volume are mined from long blocks
  - Captured by FIM but not easy to identify
- Itemsets relevant for the topic "Obama birth certificate"
  - [certificate, birth, hawaii, gov], [certificate, birth, official, hawaii, swear, obama], [abercrombi, certificate, birth, obama], ... [address, president, union, obama]

UNIVERSITY OF Waterloo

- CLEARLY specify the work in progress part
- Exponential weighting is used to give higher weight to more recent itemsets, and prevent old spikes of interest from resurfacing:

$$w = (alpha) * support(t) + (1 - alpha) * support(t-1)$$

NON-BURSTY TOPICS

- Itemsets for the week ending on 31 Jan. 2011, from the subset of Tweets related to "Obama birth certificate"

NMI	Correlation
[toothpast, radioactive],[#912, #libertarian],[abercrombi, fitch], [abercrombi, hollister]	[bioportfolio, news],[lallane, jack],[#ocra, #tcot],[#tlot, #tcot],[emanuel, ballot, rahm]

- Working on finding measures to surface informative itemsets. Temporal change of correlation is promising.
  - Example: Spreadsheet of 30 Jan 2011 in 1 hour blocks

33

TREC 2013 TEMPORAL TRACKS

- Time receives a lot of attention in next year's TREC
  - Temporal Summarization
    - Useful, new and timely updates about an *event*
  - Knowledge Base Acceleration
    - Filter a stream of content for info about an *entity*
  - Real-time search

34

CONCLUSION

- Frequent Itemset Mining (FIM) is suitable for social media
  - Short document length makes it possible
  - Support threshold filters out social content
  - Fast and scalable
- FIM captures important topics
  - Query expansion directly from itemsets increases retrieval performance

UNIVERSITY OF Waterloo

35

Query Expansion shows that itemsets is a good synopsis of the twitter stream

OPEN QUESTIONS

- In order of importance
  - Detecting topics of sustained high volume
  - Non-parametric measure of positive, negative and no correlation for NxN contingency table
  - Selecting the block and window lengths dynamically
  - Merging mining results or intermediate results

FREQUENT ITEMSET MINING IS PROMISING FOR SUMMARIZING MICROBLOG STREAMS DYNAMICALLY

Thank you! Questions?

UNIVERSITY OF Waterloo

But again, the target is to summarize the stream of publicly shared posts on social media.. we just happen to have access to a dataset of public posts from Twitter, and would be happy to experiment with data from other social networks such as Google+.

RELATED WORK

- Ritter et al (2012) use "event phrases" to extract a calendar from Twitter. http://statuscalendar.cs.washington.edu
  - Uses a POS tagger trained for this purpose.
- Ramage et al. (2012) use Supervised Latent Dirichlet Allocation (SLDA) to extract events from Twitter.
  - The runtime on one week of tweets is 4 days on a cluster of 24 machines.

UNIVERSITY OF Waterloo

38

RELATED WORK

- Petrovic et al. (2010) do first story detection on Twitter using clustering.
  - Their main contribution is using Locality Sensitive Hashing (LSH) to improve performance of clustering.
- Their system successfully decides to start a new cluster using a threshold on the cosine distance of new Tweets.
- Classifying clusters into events, neutral, and spam works well for detecting spam. Events are actual news events.

RELATED WORK

- Lin et al. (2011) studied the evolution of language models in Twitter. Two problems were investigated:
  - Different methods of keeping a history of recent Tweets
  - Different methods of smoothing the model on recent history using a background model
- Choi and Croft (2012) used temporal features to do query expansion on Twitter, based on the ReTweet (RT) feature
  - Expansion terms were selected from periods of high RT rate

UNIVERSITY OF Waterloo

• B. Huberman, D. Romero, and F. Wu. (2008). Social networks that matter: Twitter under the microscope. *First Monday, 14(1-5)*.
• J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. (KDD '09), ACM, pp. 497–506.
• I. Ounis, J. Lin, I. Soboroff. (2011). Overview of the TREC-2011 Microblog track. In *Proceedings of TREC 2011*. Gaithersburg, USA, 2011.
• R. Agrawal and R. Srikant. (1994). Fast algorithms for mining association rules. In *Proceedings of VLDB'94,* pp. 3-14.
• J. Han, J. Pei, and Y. Yin. (2000). Mining frequent patterns without candidate generation. *SIGMOD Record*, 29(2):1–12
• S. Petrov and R. McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.
• J. Lin and G. Mishne. 2012. A Study of "Churn" in Tweets and Real-Time Search Queries. In *Sixth International AAAI Conference on Weblogs and Social Media.*
• H. Li, Y. Wang, D. Zhang, M. Zhang, and E.Y. Chang. 2008. PFP: parallel FP-growth for query recommendation. In *Proceedings of the 2008 ACM conference on Recommender systems (RecSys '08)*. 107-114.
• C. H. Lau, Y.F. Li and D. Tjondronegoro. 2011. Microblog retrieval using topical features and query expansion. In *Proceedings of TREC 2011*, Gaithersburg, USA, 2011.
• V. Ganti, J. Gehrke, and R. Ramakrishnan. 2002. Mining data streams under block evolution. SIGKDD Explorer Newsletter 3, 2 (January 2002), 1-10.
• Y. Zhu and D. Shasha. 2003. Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03)*. 336-345.
• P. Liang, J. F. Roddick, A. Ceglar, A. Shillabeer, and D. de Vries. 2009. Discovering itemset interactions. In *Proceedings of the Thirty-Second Australasian Conference on Computer Science (ACSC '09)*. 133-140.
• Y. K. Lee, W.Y. Kim, Y. D. Cai and J. Han. 2003. CoMine: Efficient Mining of Correlated Patterns. In *Proc. 2003 Int. Conf. Data Mining*. 581-584.
• S. Brin, R. Motwani and C. Silverstein. 1997. Beyond Market Baskets: Generalizing Association Rules to Correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data (SIGMOD '97)*. 265-276.
• A. Ritter, Mausam, Oren Etzioni and Sam Clark. 2012. Open Domain Event Extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*.
• A. Smola and S. Narayanamurthy. 2010. An architecture for parallel topic models. In *Proceedings of VLDB Endowment* 3, 1-2 (September 2010), 703-710.
• S. Petrovic, M. Osborne, and V. Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 181-189.
• J. Lin, R. Snow, and W. Morgan. 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*
• J. Choi, and W. B. Croft. 2012. Temporal Models for Microblogs. In *Proceedings of the International Conference of Information and Knowledge Management (CIKM' 12). Maui, Hawaii, USA.*

41