View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Waterloo's Institutional Repository

Efficient Temporal Synopsis of Social Media Streams

by

Younes Abouelnagah

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2013

© Younes Abouelnagah 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Search and summarization of streaming social media, such as Twitter, requires the ongoing analysis of large volumes of data with dynamically changing characteristics. Tweets are short and repetitious – lacking context and structure – making it difficult to generate a coherent synopsis of events within a given time period. Although some established algorithms for frequent itemset analysis might provide an efficient foundation for synopsis generation, the unmodified application of standard methods produces a complex mass of rules, dominated by common language constructs and many trivial variations on topically related results. Moreover, these results are not necessarily specific to events within the time period of interest. To address these problems, we build upon the Linear time Closed itemset Mining (LCM) algorithm, which is particularly suited to the large and sparse vocabulary of tweets. LCM generates only closed itemsets, providing an immediate reduction in the number of trivial results. To reduce the impact of function words and common language constructs, we apply a filtering step that preserves these terms only when they may form part of a relevant collocation. To further reduce trivial results, we propose a novel strengthening of the closure condition of LCM to retain only those results that exceed a threshold of distinctiveness. Finally, we perform temporal ranking, based on information gain, to identify results that are particularly relevant to the time period of interest. We evaluate our work over a collection of tweets gathered in late 2012, exploring the efficiency and filtering characteristic of each processing step, both individually and collectively. Based on our experience, the resulting synopses from various time periods provide understandable and meaningful pictures of events within those periods, with potential application to tasks such as temporal summarization and query expansion for search.

Acknowledgements

I would like to thank all the people who made this possible. First and foremost, my supervisor Prof. Charles Clarke for giving me the opportunity to join such a great university, and for providing a very encouraging learning environment. Second, my family who gave me a lot of support – and the free voice chat services that kept me close to home. Last but not least, my colleagues in this journey of development with whom I worked and grew.

I would also like the thank my thesis readers, Prof. Gordon Cormack and Prof. Mark Smucker, for their time and for the valuable feedback and remarks. Special thanks to my colleague Karim Hamdan for proofreading my thesis in full, and for volunteering to do so.

Dedication

This is dedicated to the people struggling for freedom and democracy in the Arab world, while I comfortably study in beautiful Canada.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Social Media Text

Text posted on social media outlets, such as Twitter, is a good and timely source of information about real world events [72]. It also includes posts about other topics attracting the collective attention of user communities, such as Internet memes or large scale discussions [60]. Nevertheless, the collective stream from all users is overwhelmed by personal updates and non-informative chatter [39, 35]. Furthermore, the attention span given to the majority of topics is quite short [45]. Finding posts about topics of interest in a timely fashion requires an efficient mining algorithm.

To realize the benefits of social media in giving a voice to ordinary people, the algorithm should mine the content of posts without explicitly assigning weights based on popularity of users. However, the nature of text in social media poses a challenge when applying traditional text mining algorithms. Text in social media is usually short, undermining the effectiveness of within document frequency counting. It lacks context, since posts are standalone and most platforms do not allow explicit linkage. It also lacks structure and other useful formatting cues. One type of mining algorithms that can be applied to such data is frequent itemset mining.

1.2 Frequent Itemset Mining

Frequent itemset mining algorithms count the number of times each combination of "items" appear together. This is called the support of the itemset. An itemset is considered frequent if its support is above a threshold. When mining textual data, an item is usually taken to be a unigram. A set of items are considered to be appearing together if they appear within the same document or paragraph. The frequent itemset mining family of algorithms is fast and efficient, however it is not readily suited for application on text. Following are a number of challenges faced when applying frequent itemset mining to textual data. In this thesis we address those problems, adapting frequent itemset mining to social media text without degrading its efficiency:

1. The number of itemsets mined grows with the number of distinct items – which is particularly high in the text domain. In social media, the problem is complicated by the continuous creation of new tokens [53], in the form of hashtags or usernames, and due to shortening of words because of the length limit of some platforms.

2. The number of frequent itemsets is generally high. Frequent itemset mining was originally proposed as a preliminary stage for association rules mining, which sifts through the numerous itemsets and produces a smaller number of rules associating combinations of items with each other. To reduce the number of itemsets, they may be limited by setting a high support threshold. This is not possible in text mining because the frequency of most items is low, and only a few items have high frequencies; that is, frequencies of items follow a long-tailed Zipfean distribution.

3. Function words are among the few items having high frequencies. This leads to mining a large number of itemsets which are uninformative language constructs. Even if a maximum frequency threshold is set, incurring the risk that we will filter out important itemsets, many non-English constructs will be mined because the proportion of posts in English is much higher than other languages. Notice that we do not remove function words so that we allow mining itemsets made up solely of function words, and to make the mining results easier to understand at the user level.

4. There is considerable redundancy in frequent itemsets caused by trivial differences in the language used.

1.3 Motivation and Contributions

Frequent itemset mining is suitable to the dynamic nature of social media streams. It does not require prior knowledge of the distribution of items, nor does it require selecting a few items to monitor or a number of topics to mine. It is fast, efficient and scalable. It is also robust against data sparsity, and can actually exploit this sparsity for faster computation. Unlike trending topics[1] [58], the results of frequent itemset mining include itemsets that have high frequency because of sustained interest, as well as a spike of interest.

The mined itemsets provide the vocabulary associated with events and can be used as a preliminary step for search and summarization. For example, the collection of mining results from different epochs of time can be used for temporal query expansion and document expansion [20, 24]. Expansion methods can be specially developed based on the mining results, but it is also possible to use existing methods. The results from each epoch can be treated as a document, and relevant "documents" can be used to create a *temporal profile* [40] of the query or document being expanded. For summarization, frequent itemsets can provide a good foundation for summary creation. While the frequent itemsets themselves are not summaries, since they lack qualitative properties such as coherence and cohesion, the results are understandable at the user level. As we shall see in later examples, the top ranked itemsets cover a variety of open topics, and within one topic different opinions are reported as separate contrastive itemsets.

Previous work has proposed methods tailored for the use of frequent itemsets (patterns) to improve search performance. Itemsets provide better semantics than keywords, and have better statistical properties than phrases [87]. By using carefully selected itemsets and filtering out "meaningless" ones, *pattern-based* approaches achieve better performance than keyword-based approaches [88]. However, the choice of "interesting" or "informative" itemsets and filtering out "meaningless" ones remains an area of active research. In Li et

[1] http://blog.twitter.com/2010/12/to-trend-or-not-to-trend.html

al. [51] itemsets are mined from paragraphs of newswire text, and are used to determine term weights for query expansion. Improvements in performance have been achieved by using itemsets taken from a training set of related documents, as well as ones from unrelated documents. In a more recent work, an unsupervised version that relies on co-occurrence of itemsets within the same paragraph has been proposed [4].

In this thesis, we propose efficient methods for filtering out non-informative frequent itemsets, reducing redundancy in the mining results, and ranking the selection according to temporal novelty. Our methods exploit the dynamics of social media and make use of the collaborative filtering that users naturally undergo on social media, by sharing interesting posts and participating in conversations. Our main contributions are as follows:

- We propose the use of variable length term N-grams as items to avoid mining uninformative language constructs as itemsets.
- We propose a condition for selecting informative itemsets, and a clustering scheme for reducing redundancy in the mining results.
- We propose a formula for ranking itemsets according to temporal novelty.

The effectiveness of our methods is shown quantitatively, in terms of the reduction in the number of itemsets at each processing step. The quality of the final outcome is verified by showing mining results from various time periods.

1.4 Thesis Outline

The next two chapters provide necessary background. We start by discussing related work in chapter 2, then we explain the algorithm on which we build our work in chapter 3. In chapter 3, we also introduce the dataset we use, and the data prepocessing necessary for running a frequent itemset mining algorithm on social media text effectively. The most important contribution in this chapter is a simple method for flattening the Zipfean distribution of items, and thus reduce the number of itemsets made up of function words. Chapter 4 presents our main contributions for creating temporal synopses from frequent

itemsets. We present in it our proposed condition for selecting itemsets, the clustering scheme for grouping together itemsets according to their topic, and a ranking formula that is suitable for ranking clusters according to temporal novelty (based on the collective Infomration Gain of itemsets in them). In chapter 5, we empirically evaluate the efficiency and the effectiveness of our proposed methods. Finally, we conclude our presentation and suggest future directions in chapter 6.

1.5 Terminology and Notation

Classically, frequent itemset mining is applied to a *database* of *transactions* made at a retail store. This terminology is suitable for market basket data and we retain it out of convention, even though we are mining text where the terms "corpus" and "document" are normally used. Because of the dynamic nature of social media, rather than giving the whole database as input to mining algorithms, the input is an *epoch* of data; data with timestamps within a certain period of time. The epoch's *span* is the length of this period in hours, and the *volume velocity* at this epoch is the number of transactions in the epoch divided by its *span*.

We now introduce the notation used in this thesis:

- $W = \{w_1, w_2, \ldots, w_n\}$: The set of all items occurring within the epoch of data being mined. Items can be terms or term N-grams in this thesis.
- $t_a = \{w_{a_1}, \ldots, w_{a_m}\}$: A transaction made up of a set of items. Each transaction has a sequential id derived from its timestamp (denoted by the subscript letter).
- $E^{span} = \langle t_a, t_b, \ldots, t_v \rangle$: An epoch of data of a certain span, such as an hour, made up of the sequence of the transactions created within this hour.
- $s \subset W$: An itemset; any possible combination of items.
- $T_s = \{t : t \in E \, and \, s \subseteq t\}$: All transactions containing itemset s. We refer to it as the itemset's postings list, as is common in the Information Retrieval (IR) literature.

Chapter 2

Background and Related Work

2.1 Frequent Itemset Mining

Frequent itemset mining comprises a large body of work that goes back to the early 1990s [2]. We categorize frequent itemset mining algorithms into 1) algorithms that operate in the item space, 2) algorithms that operate in the transaction space, and 3) algorithms that operate in the solution space. While a comprehensive survey is beyond the scope of this thesis, we discuss at least one representative from each category of algorithms.

2.1.1 The Apriori Algorithm

Agrawal et al. [3] proposed the Apriori algorithm as part of the first efficient solution to the problem of association rules mining [2]. In the domain of market basked data, an association rule is an implication that an item is likely to be purchased in the same transaction with a certain set of items. Association rules are mined by finding itemsets that co-occur together frequently, then producing a rule with each individual item as a consequent to the purchase of the rest of the itemset. The Apriori algorithm finds frequent itemsets of increasing lengthes iteratively. It starts by counting frequencies of individual items (itemsets of length 1). Then it iteratively increases the length of itemsets. At each

iteration, it generates candidate itemsets of length k (k-itemsets) by merging frequent itemsets of length (k-1) ((k-1)-itemsets) that differ in only 1 item, usually the last item given a certain total ordering of items. By using only the frequent (k-1)-itemsets for generating candidate k-itemsets, many possible k-itemsets are implicitly pruned, based on the anti-monotonicity between itemsets' support and length: all subsets of a frequent itemset must be frequent. However, each candidate has to be explicitly checked to verify that it does not have any infrequent subsets.

Apriori and algorithms based on it suffer performance degradation and large increases in memory requirement when the number of distinct items is high. These limitations are caused by the candidate generation bottleneck. A large number of candidates can be generated, especially in early iterations of the algorithm. Consider, for example, the generation of candidate 2-itemsets from a database. This generation requires producing all unordered pairs of 1-itemsets (terms), after pruning out rare ones with frequency less than the support threshold. In many domains, including text mining, the number of frequent 1-itemsets is large enough to prohibit generating a number of candidates in the order of this number squared. In text mining, a rather low support threshold has to be used, because the frequency of terms follow a long-tailed Zipfean distribution.

The Apriori based family of algorithms operates in the item space in the sense that an itemset can be expanded by any item regardless of whether there is any support for the combination. Operating in the transaction space avoids this blind enumeration of unsupported candidates, and the subsequent DB scan required to count the support of those candidates. This involves the creation of a succinct representation of the DB that is well suited for itemset mining.

2.1.2 The FP-Growth Algorithm

The first algorithm proposed based on the idea of avoiding candidate generation was FP-Growth [30]. The database is represented as a frequency ordered prefix tree called the FP-tree. Items with low support are removed, so that any node in the FP-tree is necessarily an item with enough support. An FP-tree imposes the invariant that within each branch the frequency is non-increasing from the root down to the leaves. This increases the chance

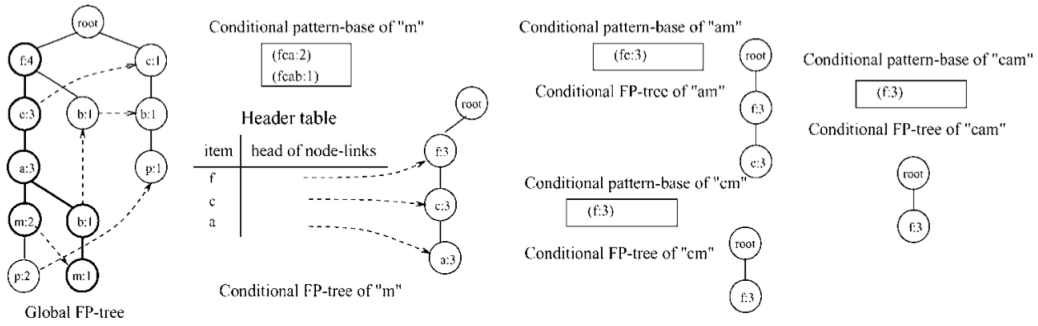


Figure 2.1: A few steps in the FP-Growth algorithm as depicted by Han et al. [31]

of finding common prefixes. An FP-tree also contains links between nodes representing the same item. Those links are used to create the *conditional* FP-tree for an item, which is the FP-tree made only from branches containing this item, excluding any items with lower frequency. The algorithm proceeds by iterating over items in increasing order of frequency. Each item is first output as an itemset, and then its *conditional* FP-Tree is created and recursively mined. At each recursive call, the current item is appended to a growing prefix, which is prepended to all itemsets mined from the conditional FP-tree. Figure 2.1, borrowed from Han et. al [31], shows a few steps in the FP-Growth algorithm.

The navigation of a representation of the database avoids the candidate generation bottleneck, by directly deriving itemsets from the transaction space. However, the creation of such representation is a bottleneck itself. Building conditional pattern bases and sub-FP-trees becomes time and space consuming as the recursion goes deep and the number of patterns becomes large [85].

2.1.3 Solution Space Algorithms

The last category of frequent itemset mining algorithms comprises algorithms that operate in the solution space. It can be argued that algorithms that traverse a representation of the DB are operating in the solution space rather than the transaction space – given that they avoid generating candidates with low support. However, in our categorization

we focus on the logic behind how itemsets are generated or pruned rather than the data structure used to generate them. After all, any algorithm must calculate the support of itemsets, and this requires the algorithm to either maintain a representation of the DB or scan the transactions. Algorithms that operate in the solution space prune itemsets that do not possess certain properties. The limited number of itemsets that possess the desired property are sufficient to deduce the rest of the solution (other frequent itemsets and their support information). The reduction in the number of itemsets mined improves the performance of such algorithms, even if they rely on candidate generation to enumerate possible itemsets before checking for the desired property. It is not always necessary to produce the full solution since the properties usually filter out only redundant itemsets, and in many cases the elect itemsets can be used directly.

The *closure property* [65, 66, 96] prunes an itemset if it has the same support as its supersets. It is easy to see that all itemsets and their support can be derived from the set of *closed* itemsets. A formal proof is given by Pasquier et al. [65]. The smallest itemset of an equivalence class of itemsets having the same support is called a *generator* [44] or a *free set* [12]. It is also possible to keep only the generator of each equivalence class, but in this case some infrequent itemsets has to be kept in order to be able to calculate the support of all frequent itemsets [44]. A different selection of itemsets can be made according to the *non-derivable* property [17, 18], which is linked to the closed property.

Finding closed itemsets can be done by arranging possible itemsets into a Galois lattice, then traversing the lattice using the Galois connection between itemsets and transactions in which they appear. Figure 2.2 shows an example of a Galois lattice. The Galois connection is a pair of operators, one to map an itemset to transactions in which it appears, and another to map a set of transactions to the largest itemset that appears in all of them. Applying the two operators in cascade grows an itemset directly to its closed superset. More details can be found in Pasquier et al. [66] and Zaki et al. [96].

Some algorithms take advantage of the closure property to reduce the search space more dramatically, such as CLOSET [67] and LCM [81]. Our work is based on LCM (which stands for Linear-time Closed itemset Mining) [81]. As a starting point for our work, we use the implementation of LCM submitted to the workshop for Frequent Itemset Mining Implementations (FIMI) in 2004 [41], which was the workshop's award winner. The

TID	Items
1	A B C
2	B C E
3	A B C E
4	B E
5	A B C E

The transaction database

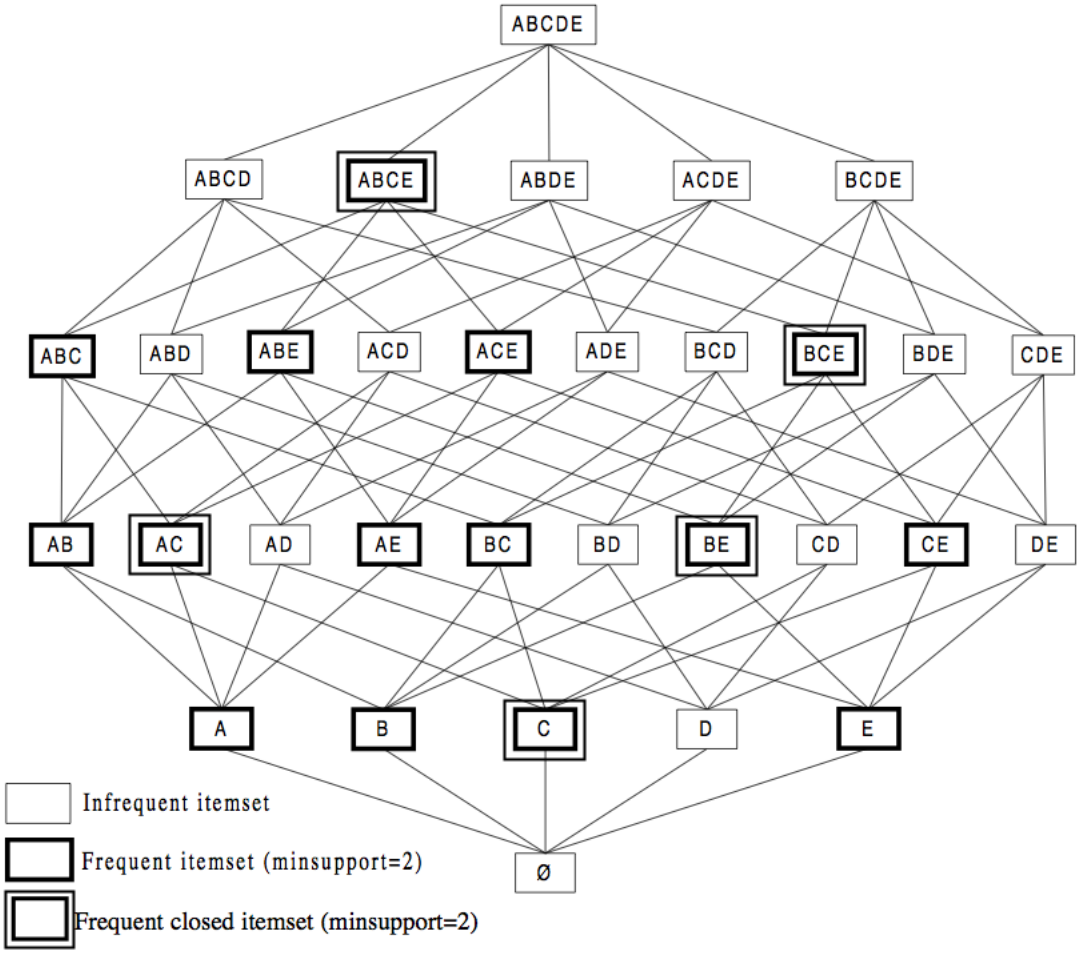


Figure 2.2: Example of a Galois lattice adapted from Pasquier et al. [66]

LCM algorithm is robust against data sparsity, and has low memory requirements since it does not store intermediate results. We will explain the closure property and describe LCM in detail in section 3.1. We proceed by discussing other areas of related work.

2.2 Election of Representative or Interesting Itemsets

The problem of having many itemsets, with redundancy and noise, can be addressed by picking out representative itemsets that satisfy a certain condition. The representative itemsets are not necessarily sufficient to deduce the rest of the solution (itemsets and their support information). In this thesis we propose a condition for selecting interesting itemsets, which we call *distinct* itemsets. Related approaches to picking out representative itemsets are 1) choosing a set that will minimize the reconstruction error of the full solution, and 2) limiting the number of itemsets to a user specified number.

The δ-free [13], δ-covered [90], and *master* itemsets [92] are examples of conditions motivated by providing a compressed representation of itemsets through sacrificing support information. The δ-free condition selects an itemsets if its support is different from the support of all its subsets by at least δ. This selection of itemsets can be used to approximate the support of other frequent itemsets with a guarantee on the error, but only if some infrequent itemsets are included in the selection. The δ-free condition implies setting an upper bound on the strength of association rules that can be derived from selected itemsets. It is suitable for compressing mining results from dense datasets, but it is exactly the opposite of *distinctiveness* condition we propose in this thesis.

The δ-covered condition is also the opposite of the *distinctiveness* condition, as it "relaxes the closure condition to further reduce pattern set size" [55]. On the other hand, the *distinct* condition strengthens the closure condition. The δ-covered condition sets an upper bound on the confidence of the rule that an itemset implies a superset; if an itemset implies its superset with confidence greater than δ, then it is considered to be covered by the superset and can be pruned. On the other hand, the *distinct* condition sets a lower bound on the confidence of a derived association rule. Nevertheless, the clustering scheme Xin et al. [90] propose (δ-clusters) is similar to the *strongly closed itemset* clusters

proposed in this thesis. The efficient version of the clustering algorithm, RPLocal, limits the search space for clustering candidates using a similar reasoning to the reasoning used in our algorithm – exploiting the sequence in which frequent itemsets are found. The main difference between δ-clusters and *strongly closed itemset* clusters is that δ-clusters are soft clusters, while *strongly closed itemset* clusters are hard clusters. Also, the distance measure used in δ-clusters is the Jaccard distance, while the *strongly closed itemset* clusters are based on association rule confidence.

Yan et al. [92] proposed a method to choose k itemsets as representatives of the itemsets in the full solution, which they call *master* itemsets. Rather than leaving k as a user specified parameter, a method is proposed for choosing k by minimizing the restoration error of support information. A *master* pattern is the union of all itemsets in a cluster, similar to our proposed *strongly closed itemsets*. While the use of clustering is similarly motivated by the trivial difference between itemsets, the k *master* itemsets have to cover all itemsets. On the other hand, *strongly closed itemsets* actually avoid clustering together itemsets representing different topics or different opinions within a topic. Furthermore, Yan et al. improve performance by approximating the calculation of the distance between itemsets, to avoid accessing the data. We can achieve high performance without approximation because our mining results include the postings list of each itemset.

Another approach to choosing itemsets is according to their interestingness. The interestingness of itemsets (patterns) has been an area of active research for a long time. According to Silberschatz and Tuzhilin [75] interestingness of an itemset can be *objective* or *subjective*. Commonly used *objective* measures include (a more extensive coverage is provided by Geng et al. [26] and Tan et al. [77]):

1. *Support*: The number (or fraction) of transactions that contain the itemset.

2. *Confidence*: The confidence of a rule derived from the itemset. A rule is derived from the itemset by treating one item, w, as the consequent of the rule while the rest of items, $s \setminus w$, are the antecedent. The confidence of the rule is then defined as:

$$conf(s \setminus w \rightarrow s) = \frac{|T_s|}{|T_{s \setminus w}|} \tag{2.1}$$

Since different rules with varying confidence can be derived from an itemset, either the minimum (*all-confidence*) or the maximum (*any-confidence*) is used [63].

3. *Lift* [16]: The probability (support) of the itemset over the product of the probabilities of all items in the itemset. This is a measure of dependence and can be linked to Pointwise Mutual Information [14]. Such a measure is suitable for knowledge discovery or exploratory search tasks, where the goal is to extract correlated itemsets or significant phrases.

4. *Bond (or Coherence)* [63]: The support of an itemset over the number of transactions containing *any* of its items.

An important property of a measure, upon which all Apriori based mining algorithms are built, is the downward closure property; the value of a downward closed measure cannot increase as the size of the itemset increases. All the above measures are downward closed [9] with the exception of *lift* which is upward closed (cannot decrease with the increase of the size of the itemset), and *any-confidence* which is neither upward nor downward closed. Algorithms that mine itemsets based on the above measures were proposed [47, 42], but if an application specific measure is to be used then it has to be evaluated in a post mining stage or within the Redundancy-Aware Top-K Framework [89].

The Redundancy-Aware Top-K Framework [89] selects the top k itemsets according to an interestingness criterion. The major difference between their approach and all of the previous work is that it emphasizes both interestingness and redundancy when selecting the top k itemsets. The itemset interestingness is defined in the context of the application; summarizing the whole collection of itemsets is not the goal. Our work fits in this framework: we reduce redundancy through filtering and clustering, and we propose a ranking that orders itemsets according to some definition of interestingness. If only the top k itemsets are desired, itemsets ranked at positions higher than k can be omitted. An explicit number of itemsets needs not be set, however. This makes our computational model more efficient, since we do not need to solve a constrained optimization problem. In their experiments, Xin et al. [89] used two application specific interestingness measures: they used one that is based on TF-IDF for extracting interesting itemsets from a text corpus,

and another that is tailored for predicting block fetches in storage systems [52]. We rank itemsets according to their novelty using a measure based on Information Gain. Preliminary experimentation with various other measures of interestingness ruled them out as ineffective for our purposes.

Subjective measures of interestingness select itemsets that are actionable or unexpected [76]. Selecting unexpected itemsets entails building a probabilistic model to use it for estimating itemsets probabilities. Jaroszewicz and Simovici [38] use a Baysian network as a model, which is suitable for cases where items denote discrete values of a limited number of attributes. A commonly used model is the maximum entropy model [83, 78, 57]. The Maximally informaTiVe itemsets (MTV) algorithm [57] uses a greedy algorithm to select k itemsets whose probabilities diverge the most from the estimate given by a maximum entropy model. The maximum entropy model is initialized as a uniform distribution. After selecting each of the k itemsets it is updated to fit the selected itemsets actual frequencies using the *iterative scaling* procedure [23]. If k is set to infinity, the algorithm will generate the best model according to the Bayesian Information Criterion (BIC). Since BIC incorporates a penalty term for the number of parameters (itemsets included in the model), this selection will minimize redundancy while maximizing surprise (divergence from model). The model can be initialized using itemset frequencies known from prior knowledge. It is also possible to use the Minimum Description Length (MDL) as a criterion for selecting itemsets. We compare our selection of itemsets to ones obtained from MTV using the implementation provided by the authors[1]. However, the value of k has to be kept low (at most 100) for the algorithm to finish without an error, even though it is running without a limit on its resource usage on a machine with 256 GB of RAM. In our work we focus on efficiency, allowing the summary to include larger numbers of itemsets than the values of k for which MTV finishes in a reasonable time (at most 10). This is crucial for scalability to the volumes of data typical to social media streams.

Another approach to picking out itemsets is choosing ones that can be used to compress the data (not the itemsets). The KRIMP algorithm [82] is a good example of methods that follow this approach. Our goal is different because we aim to filter out itemsets pertaining to personal updates, which make up a large portion of social media data.

[1] http://www.adrem.ua.ac.be/implementations

2.3 Frequent Itemsets in Search and Summarization

Different ways to use frequent itemset mining in search have been proposed. Possas et al. [70] proposed using closed itemsets in place of terms, using them in exactly the same way that terms are used. In the tokenization process, they use CHARM [96] to mine closed itemsets of terms from each document. They then add the document to the postings lists of each closed itemset it contains. Queries are tokenized similarly and results are ranked using TF-IDF. This simple approach achieves increased precision on various test collections, possibly because frequent itemsets take into account collocation information. The approach is simplistic, however. The overlap between itemsets was not taken into account while tokenizing documents and queries into itemsets. Also, the mining algorithm was not modified to accommodate the high number of terms and the greater length of documents compared to the length of transactions typical to market basket data. Finally, the number of itemsets is higher than the number of terms so the index size will increase.

Wu et al. [88, 87] proposed the use of sequential closed patterns in place of terms, in what they call a *pattern based approach* to information retrieval. They consider patterns as surrogates to phrases that have better statistical properties than phrases, thus achieving a break through in the performance of *phrase based approaches*. "Although phrases contain less ambiguous and narrower meanings than individual words, the likely reasons for the discouraging performance [50] from the use of phrases are: (1) phrases have inferior statistical properties to words, (2) they have low frequency of occurrence, and (3) there are a large number of redundant and noisy phrases among them." [87] They addressed the problem of overlap between itemsets, and they developed an algorithm specifically tailored for mining closed sequential patterns from text. Documents are broken into paragraphs such that a paragraph is the unit of mining and retrieval. This alleviates the problem that documents are usually much longer than traditional market basket transactions. Albathan et al. [4] extended the work of Wu et al. by proposing a method for further reducing the number of closed patterns. Lau et al. [46] used related methods for term weighting in pseudo-relevance feedback for Twitter search, and achieved substantial improvements over a baseline.

In an earlier work [1], we have also explored the viability of using frequent itemsets for query expansion in microblog search. We could achieve improvements over a strong baseline on the TREC 2011 microblog track queries. As a baseline, we used Okapi BM25 [73] with its parameters tuned to achieve the best performance on short documents. The baseline's performance ranked high compared to the TREC 2011 submissions. After query expansion, the performance was better than all of the TREC 2011 submissions. Queries were expanded using terms from itemsets that are most relevant to the query, according to the same relevance ranking formula used in the baseline. We used frequent itemsets mined by an unmodified mining algorithm that returns the top k itemsets according to their support[2]. However, query expansion achieved no improvement over the baseline on the TREC 2012 microblog track queries. This indicates that there was not enough relevant itemsets, because of the use of support to limit the number of itemsets. The methods proposed in this thesis can be used for mining a short ranked list of relevant itemsets, which can then be used by methods such as our earlier work, or the work of Wu et al.

This work also complements the work done by Yang et al. [93] for using frequent itemsets as a temporal summary. They have proposed a framework for storing mining results of temporally consecutive intervals (batches of data) using a pyramidal time window. Their framework allows for fast execution of temporal queries, and for tracking the evolution of topics. However, the itemsets stored for each interval are not selected for providing the best summary. Instead, they are selected according to their utility in compressing the data. The temporal summary is created from transactions relevant to a query in a separate step. First, the pyramidal time window is queried using keywords and a specific time intervals (for example, "world cup" on the day before a certain match). Then, non-negative matrix factorization is used to extract topics from the returned transactions. The pyramidal time window structure can be used to store itemsets selected by our methods, thus directly providing a temporal summary without the need for the matrix factorization step. Our methods exploit the nature of social media to choose topically relevant itemsets without the need for a query. Social media has been shown to be a good source of real-time information, as we will discuss in the next section.

[2] https://cwiki.apache.org/MAHOUT/parallel-frequent-pattern-mining.html

2.4 Social Media as a Source of Information

The problem of finding information in social media posts is motivated by the uniqueness and variety of information that can be found in these media. Some stories are reported mainly on social media because of censorship on journalism, such as the 2009 Iranian elections protests (nicknamed "Iran's Twitter Revolution"[3]). Other stories are reported mainly on social media because the content creator happens to be at the right place at the right time. This could be by chance, like in the case of the emergency landing of a US Airways plane on the Hudson river in New York[4]. This could also be an intentional act of citizen journalism. Citizen journalists are "the people formerly known as the audience," who "were on the receiving end of a media system that ran one way, in a broadcasting pattern, with high entry fees and a few firms competing to speak very loudly while the rest of the population listened in isolation from one another – and who today are not in a situation like that at all."[5]

In a study of the use of Twitter as a news medium during the 2011 Tunisia and Egypt revolutions, Lotan et al. [56] categorized 963 accounts that made influential posts according to the *type of actor*. A post is considered influential if it is among the top 10% of posts in terms of the number of near duplicates made out of it; posts that had at least 16 near duplicates in the Tunisia data, or at least 19 near duplicates in the Egypt data. The types of actors included *mainstream media organizations* and *mainstream new media organizations*, which are news and media organizations with and without offline presence, respectively. It also included *journalists*, who are individuals employed by media organizations or who regularly work as freelancers for them.

In both datasets, media organizations make up about 11% only of the 963 accounts (approximately 7% for ones with offline presence and 4% for ones that exist solely online). Journalists make up another 15%, so the total of mainstream media affiliated accounts is approximately 26%. On the other hand, more than half of the accounts from which influen-

[3]http://www.theatlantic.com/technology/archive/2010/06/evaluating-irans-twitter-revolution/58337/

[4]http://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html

[5]http://archive.pressthink.org/2006/06/27/ppl_frmr.html#more

tial tweets originated belong to individuals not affiliated to the government or mainstream media (55% for Tunisia and 56% for Egypt). Out of those individuals, almost half are ordinary people who do not self identify and cannot be identified by assessors as bloggers or activists.

The rate at which an account makes posts is approximately 16 per hour for mainstream media affiliated accounts, and approximately 10 for individuals accounts (thus a mainstream media account makes 60% more posts). Also the probability that a post will be echoed is 0.88 for mainstream media accounts, 0.55 for journalists, 0.4 for bloggers and activists, and 0.31 for individuals (this is the probability that any follower will echo the post, the average number of followers of an account belonging to an ordinary person is much lower than that of an account belonging to an actor of any of the other types). However, collectively the fraction of highly echoed threads originating from ordinary people is on a par with the fraction of threads originating from a journalist. Threads originating from ordinary people, bloggers and activists together make up more than half of highly echoed threads, followed by threads originating from journalists (around 20%), then threads originating from mainstream media (around 10%, divided equally between media with and without offline presence).

All this emphasizes the influence citizen journalists and ordinary people now have on the information creation and dissemination process. User created content is rich with unique content tackling a variety of topics from different points of view. However, it is dominated by non-informative chatter and personal updates. In a study comparing the use of Twitter's search feature to Web search, Teevan et al. [79] found that Twitter is mostly used for getting timely information: updates about events, news about topics of interest (such as technology news), and realtime local updates (such as traffic jams). It was noted that Twitter search could be the main way of accessing Twitter, since posts from the followed network could to be outside of the user's topical interests. This is supported by findings from an earlier study by Naaman et al. [61], in which 3,379 tweets from 350 accounts belonging to individuals were studied. The tweets were hand coded into different categories, then the activity of users in terms of what type of tweets they make was analyzed. It was found that 40% of the studied tweets are personal updates (about "*me now*"), and that users can be clustered into a big cluster of "*meformers*" (80% of

the users, half of their posts are personal updates) and a small cluster of *informers* (20% of the users, half of their posts are for sharing information). The study also found that most people engage in some form of *meformer* activity; on average, users had 41% of their messages in the "*me now*" category. Therefore, filtering out posts by *meformers*, or giving higher weights to posts by *informers* would not be sufficient for filtering posts about "*me now*" and finding informative posts made by ordinary people.

To tap into social media as a source of information, social media users collaboratively select good content. On Twitter, users can *retweet* an interesting post, which means forwarding it to their followers. The dynamics of retweeting was studied by Boyd et al [15]. One of the findings was that posts being retweeted are frequently modified to accommodate additional text, or at least the notation meant to indicate that the post is a retweet. The limit on tweets length might be the reason behind this modification, but it can also be a way to emphasize what the retweeter finds most interesting. The selection of what to keep from the post is actually a form of collaborative filtering itself, because the most interesting parts has to be retained. Parts that are selected by many people are likely to be good summaries of what is interesting. Using frequent itemset mining, the words in those selections would be mined as itemsets. In chapter 4 we discuss how to exploit such social media dynamics to create temporal synopses. In the next section, we discuss other methods for finding emerging topics in social media.

2.5 Topic Detection and Tracking for Social Media

The problem of Topic Detection and Tracking (TDT) was introduced by Allan et al. in 1998 [6]. Topic detection is the discovery of previously unknown topics, where a topic was first defined as an "event" then its definition was broadened to include the triggering event as well as other events and activities directly related to it. Our work can be categorized as a method for topic detection from social media, or more specifically "event" detection since we make no effort to group events into topics. Ever since it was introduced, TDT has been an area of active research. Besides the fact that there is always room for improvement, TDT was originally targeting the domain of newswire data, and the solutions developed

for this domain are not directly applicable to data from other domains.

In the domain of social media, Petrović et al. [68] used an adaptation of Locality Sensitive Hashing (LSH) [36] to perform first story detection at scale. The computation framework is the same as the one used in traditional systems [94, 7]:

- The distance between a tweet and its nearest neighbour from the preceding tweets is calculated.
- If the distance is less than a threshold, then the tweet is added to the topic represented by the cluster containing the nearest neighbour.
- If the distance is greater than the threshold, a new cluster is created. The new cluster represents a new topic, with the current tweet as its first story.

The adaptation of LSH that Petrović et al. propose reduces the number of times a tweet is considered far from all of its neighbours while it actually is not. In case of failure to find a near neighbour, a tweet is compared with the last 1000 tweets, updating the minimum distance if necessary. The performance of the modified LSH on a traditional TDT task (TDT-5) was comparable to the UMass FSD system [7]. On the other hand, the run times of the two systems are 2 hours and 28 hours respectively. Another experiment was done on a dataset of a 160 million tweets collected over a period of 6 months from Twitter. The UMass system failed to terminate, while the LSH based system could process each tweet in bounded space and time. One problem remains, however. Their system does not achieve high precision in determining if a tweet cluster pertains to an event or not. Actually, their system can detect spam clusters with high precision ($average\,precision = 0.963$), but not event clusters ($average\,precision = 0.34$). The problem of identifying event clusters was further studied by Beker et al. [10], where they used a Support Vector Machine (SVM) classifier and achieved a higher precision. However, the test set used was limited to 5 hours of Twitter data and the precision decreases as the number of clusters increases.

Another approach to event detection follows the framework of bursty keywords detection proposed by Kleinberg [43]. In this framework, a set of keywords are tracked – possibly all tokens in the dataset. A state machine is used to indicate the state of each keyword,

where consequitive states correspond to higher levels of *burst* intensity. Each state transition is associated with a cost, and a keyword is moved between states such that the overall cost of its transitions is minimized. A simplified version of this framework is used by Parikh and Karlapalem [64]. In their work, there are only two states for each token: the normal state, and the bursty state. A token is moved to the state of being bursty if the increase in its frequency exceeds a threshold, and it stays in this state for a fixed time interval. The timeline is divided into fixed length intervals, and a token is totally ignored in intervals when it is in the normal state. Tokens are then clustered according to the similarity of the sets of bigrams with which the co-occur, and the pattern of their *appearance* (moving to the bursty state). The clusters are considered to be events, and they are ranked according to their sizes. The clustering they propose has to process data from a long interval, so that the similarity of appearance can be calculated. This approach is not suitable for stream processing. Besides, the number of clusters produced by their methods is small compared to the number of events one would expect to have happened in the interval of time processed (23 in 20 days and 15 in a week).

Another method that tracks the frequency of all tokens is EDCoW [86], where the wavelet transformation is used to convert counts of occurrences from the time domain to the frequency domain then auto-correlation is used to detect burstiness. However, the use of wavelet transformation is not justified specially for the enormous number of tokens in case of social media. Other methods based on burst detection track a small set of keywords. Lehmann et al. [48] track hashtags and study their usage pattern. Chakrabarti and Punera [19] track tokens appearing in tweets tagged by the name of an American football team, and use them to create a summary of a game.

Our work can also be seen in the light of meme-tracking [49], where a particularly interesting quotation (meme) is tracked across news and blog posts. The quotation being tracked is not quoted verbatim in all posts, but rather parts of it are selected and it could be slightly paraphrased. This is similar to modifications made during retweeting an interesting tweet. Leskovec et al. [49] tracked quotations made by American politicians during the 2008 U.S. presidential elections. Phrases made up of 4 words or more are tracked if they got repeated 10 times or more in blogs and news articles, where at most 25% of the repetitions could be on the same domain name. A directed acyclic graph is

constructed where vertices represent phrases and edges are weighted according to the edit distance between the phrases, as well as the number of occurrences of the longer phrase. The graph is then divided into clusters where each cluster contains one *root* that has no outgoing edges. Each vertex is assigned to the cluster with which it has the highest number of edges. Clusters are then ranked according to the number of news articles and blog posts containing any phrase from the cluster. Leskovec et al. then use the clusters and their associated news articles and blog posts to perform an analysis of the temporal variation of the volume of posts about each meme. They propose a model that fits the data well, and study the difference between blog posts and news articles in terms of how fast a meme is picked and how long it remains in focus before getting dropped. In our work, we are also using the volume of posts containing phrases and their variations (itemsets). We also do clustering but in a different way, and we work in a totally different scale of time.

O'Connor et al. [62] also start from phrases to summarize tweets about a user specified query. They select *significant* phrases which are statistically unlikely, then cluster them. Frequent itemsets are a better starting point than phrases, since there are much more phrases than frequent itemsets.

Event detection in Twitter can also be done using natural language processing to identify *event* parts of speech [72], or to identify action phrases [69]. Of course, it is also possible to apply topic modelling by Latent Dirichlet Allocation (LDA) [33] or an online varient [32, 95]. Our method tackles the problem in a totally different way than these two approaches.

Chapter 3

Mining Social Media Text

3.1 The Linear-Time Closed Itemset Mining Algorithm

In section 2.1.3 we have discussed frequent itemset mining algorithms that overcome the bottleneck of *candidate generation* by limiting the solution to itemsets with a certain property, reducing the size of the solution and possibly providing a way to quickly navigate the solution space. In this thesis, we expand on the Linear-time Closed itemset Mining (LCM) algorithm [81], which mines only *closed itemsets* [65]. The LCM algorithm is distinctively fast because it also takes hints from the transaction space during candidate generation. This also makes it resilient against data sparsity, since it will not consider a candidate that never appears in the data. The ability to efficiently mine sparse data makes LCM particularly suitable for mining social media text. In this section we explain in detail how LCM works.

Since LCM makes use of the properties of closed itemsets, we begin our presentation by discussing these properties. Informally, a closed itemset contains any item that is present in all the transactions containing this itemset. A formal definition of closed itemsets is given in equation 3.1:

$$\mathcal{C} = \{s_c : \ s_c \subset W \, and \nexists \, s_d \, where \, s_c \subset s_d \, and \, |T_{s_c}| = |T_{s_d}|\} \tag{3.1}$$

The properties of closed itemsets are as follows:

1. Adding an item to a closed itemset reduces its support.
2. A subset of a closed itemset is not necessarily closed, but one or more closed subset must exist for any itemset (formally this could be the empty set, given that any item that appears in all transactions is removed in a preprocessing step).
3. If a closed k-itemset can be extended any further then one of its supersets will be closed, however not necessarily a (k+1) superset. Itemsets that cannot be extended any further are called *maximal itemsets*, which form a subclass of closed itemsets.

Besides being much smaller than the solution space of frequent itemsets, the solution space of closed itemsets can be navigated efficiently. By using an arbitrary total ordering of items, any closed itemset can be considered an extension of exactly one of its subsets. Thus, only this subset is extended during candidate generation. All other subsets do not need to be extended by items that would lead to the longer closed itemset. This property is called *prefix preserving closure extension (PPC-Extension)* and it was proposed and formally proved by Uno et al. [81].

PPC-Extension is achieved by following three rules, which we state after a few definitions to facilitate their statement. First, an item is *larger/smaller* than another item if it comes later/earlier in the total ordering. This terminology comes from the fact that LCM is most efficient if the items are ordered in ascending order of their frequency. Second, the *suffix* of an itemset is one or more items which have to be removed to get an itemset with greater support. Notice that they are necessarily the last items added to the itemset, regardless of the total ordering. Finally, we call the first item added to the suffix of the itemset its *suffix head*. With this terminology, the rules for *PPC-Extentsion* are as follows:

1. An itemset must be extended by every item that occurs in $T_{itemset}$, except items which are *smaller* than its *suffix head*; extending by *smaller* items will lead to closed

itemsets already generated in an earlier step. Extension items are added to the itemset in turn in the order given by the total ordering, recursively applying PPC-Extension to the extended itemset before adding the next extension item.

2. After adding an extension item, w_e, to an itemset, s, we add all other items that appear in all transaction containing $s \cup \{w_e\}$ – that is, items whose frequency within $T_{s \cup \{w_e\}}$ is equal to $|T_{s \cup \{w_e\}}|$. The newly added items become the *suffix*.

3. If all items in the *suffix* are *larger* than the *suffix head* then add the itemset to the solution. Otherwise, prune this solution branch; all closed itemsets within this branch have already been generated, when processing the smallest suffix member.

Table 3.1 is an example of how *PPC-Extentsion* is used to generate closed itemsets starting from the *1*-itemset {'barack'}. The upper table enumerates $T_{\{\text{'barack'}\}}$. The lower table shows steps of itemsets generation. The current itemset along with its frequency is in column 2. Itemsets marked by an (*) are the closed itemsets that are part of the solution. The suffix of the itemset is shown in *italic*. All possible extension items and their frequencies are in column 3. Extension items that are *smaller* than the *suffix head* are shown with a line striked through them. For each itemset, the extension items are kept so that it is known which extension item is next in turn to be added. An item is bolded when its turn to be added has come. After adding each item, a pass is done on $T_{itemset}$ to enumerate and count possible extension items. To enforce a support threshold infrequent extension items would be removed after counting, but in this example there is no such threshold. Finally, column 4 is a comment explaining each step.

Table 3.1 shows that the number of steps is linear in the number of closed itemsets, and the only additional storage required, besides storage for the documents, is that required for possible extension items. Of course, this is a simplified example, but it shows in essence how LCM achieves its low run time and memory requirements. We refer the interested reader to Uno et al. [81] for a theoretical proof that the algorithm runs in linear time in the number of closed itemsets, and that this number is quadratic in the number of transactions. Performance on a real dataset is shown in section 3.2. We proceed by describing how to implement this algorithm using an inverted index.

Doc. Id	Document	Doc. Id	Document
a	barack & mitt	b	barack obama & mitt romney
c	barack obama & romney	d	barack obama

Documents (two per row)

Step	Current Itemset	Possible Extension Items	Comments
1	{barack} (4)	mitt (2), obama (3), romney (2)	Items in $T_{\{\text{'}barack\text{'}\}}$ are counted. There is not an item appearing in all transactions.
2	{*barack*} (4)*	mitt (2), obama (3), romney (2)	Rule 3: suffix is ordered, add itemset to solution. This is not shown as a separate step in the rest of the example.
3	{*barack*} (4)	**mitt** (2), obama (3), romney (2)	Items are ordered lexicographically. Adding 'mitt' to itemset.
4	{barack, *mitt*} (2)*	**obama** (1), romney (1)	Extension items reenumerated & counted.
5	{barack, *mitt*, obama} (1)	romney (1)	Rule 2: 'romney' appears in all $T_{itemset}$.
6	{barack, mitt, *obama, romney*}(1)*		Rule 3: 'obama' is the *suffix head*.
7	{*barack*} (4)	mitt (2), **obama** (3), romney (2)	Nothing more to add, back to {'barack'}.
8	{barack, *obama*} (3)*	~~mitt~~ (1), **romney** (2)	Rule 1: skipping 'mitt', adding 'romney'
9	{barack, obama, *romney*} (2)*	~~mitt~~ (1)	Rule 1: Nothing more to add.
10	{*barack*} (4)	mitt (2), obama (3), **romney** (2)	Back to {'barack'}, adding 'romney'.
11	{*barack*, romney} (2)	mitt (1), obama (2)	Rule 2: add 'obama' after 'romney'.
12	{barack, *romney, obama*} (2)	~~mitt~~ (1)	Rule 3: suffix is not ordered, prune.
13	{*barack*} (4)	mitt (2), obama (3), romney (2)	Back to {'barack'}, all possible extension items were added. Done.

Closed itemsets containing 'barack'

Table 3.1: Generation of closed itemsets by Prefix Preserving Closure Extension

3.1.1 Implementation Details

We show in algorithm 1 how to implement LCM and PPC-Extension using an inverted index. The algorithm assumes the presence of a search infrastructure, with the ability of performing boolean queries efficiently. Besides the high likelihood that such an infrastructure already exists, it is also a good representation of textual data. Other representations that are designed for databases in general might not be well suited for textual data. For example, the memory requirements of the FP-tree suffers from the sparsity of data, since the data structure is succinct only if it can find common prefixes within the constraints of its invariant. The possibility of using an inverted index in the implementation of LCM is yet another reason why it is well suited for textual data.

The algorithm takes as input an epoch of data and a support threshold. It outputs the closed itemsets with support above the threshold. Along with each itemset in the solution, it also outputs the transactions in which it occurs – which is represented as $\langle items, T_{itemset} \rangle$. The symbol $\succ$ denotes that the lefthand side succeeds the righthand side in the total ordering. In the implementation shown, it is not necessary that the inverted index's tokens list (X.tokens) follow the total ordering; all itemsets of length 1 will be considered anyway. Each 1-itemset is then expanded according to the PPC-Extension rules. Lines 11-18 are the enumeration and occurrence counting of possible extension items, except line 15 which forms the closed itemset according to PPC-Extension rule number 2. Lines 19-20 prune a solution branch according to rule number 3, and line 21 adds the itemset to the solution according to the rule's complement. Lines 22-27 extend the itemset according to rule number 1, where line 23 (along with line 5) enforces the support threshold.

The algorithm lends itself to distributed implementations. For example, a map/reduce implementation is straightforward since the only operations are counting (line 15) and projection (line 24). However, the fast execution time and the low memory requirements of the algorithm makes it possible that a distributed implementation will cause unnecessary overhead for all but the largest datasets. We experimented with Hadoop[1] and the overhead was in the order of minutes. Besides, many problems start to arise when the programming model is complicated, and Hadoop was particularly problematic.

[1] http://hadoop.apache.org

Input: *support*: Support threshold
Data: E: Epoch of data
Result: C: Closed itemsets occurring in at least *support* records

```
1  C ← {⟨∅, E⟩} ;      // ∅ is a closed itemset.  This is skipped in practice
2  X ← Inverted index of E;
3  foreach w ∈ X.tokens do
4  |   T_{w} ← X.postingsList[w];
5  |   if |T_{w}| ≥ support then // Support threshold enforcement on 1-itemsets
6  |   |   LCM({w}, w, T_{w}) ;
7  end
8  return C;
9  Function LCM(s: Current itemset, w_sh: Suffix head,
10                 T_s: Transactions (tweets) containing s) is
11 |   frequency[1 ... w_n] ← 0;
12 |   suffix ← {w_sh};
13 |   foreach t ∈ T_s do
14 |   |   foreach w ∈ t do
15 |   |   |   frequency[w]++;
16 |   |   |   if frequency[w] = |T_s| then  suffix ← suffix ∪ {w} ;
17 |   |   end
18 |   end
19 |   if ∃v ∈ suffix : w_sh ≻ v then
20 |   |   return ;                  // Prune according to PPC-Extension Rule 3
21 |   C ← C ∪ {⟨s ∪ suffix, T_s⟩};
22 |   foreach v ≻ w_sh and v ∉ suffix do
23 |   |   if frequency[v] ≥ support then
24 |   |   |   T_{s∪{v}} ← T_s ∩ T_{v} ;                  // Results of query s AND v
25 |   |   |   LCM(s ∪ suffix ∪ {v}, v, T_{s∪{v}})
26 |   |   end
27 |   end
28 end
```

Algorithm 1: The LCM frequent itemset mining algorithm

3.2 Using LCM for Mining Social Media Text

3.2.1 Dataset

Throughout this paper we use data collected from the Twitter public stream[2] since October 1st, 2012. We use only tweets written in the Latin script to facilitate tokenization using white space and other word boundaries. We collect only the tweet text to avoid reliance on any features specific to a certain social medium, and to make the algorithms applicable to other media where text is short such as Facebook or Google+ status updates. YouTube comment data is a particularly promising candidate since comments are limited to 500 characters; the itemsets could provide textual highlights of non-textual content.

We remove duplicate original tweets (not retweets) using a Bloom filter [59]. This filtering removes spam tweets sent by botnets, averaging at 2.86% of the stream. There were two disruptions during data gathering. One in late October because of Hurricane Sandy which made the Twitter service unaccessible. The other was in late December because of technical problems on the server gathering the data.

The average number of tweets per hour (*volume velocity*) in the data is 119,035.49 tweets (n = 2,546 hours, standard error = 491), and the average number of tweets per day is 2,705,931.83 tweets (n = 112 days, standard error = 55,155). These are simple moving averages calculated by summing the number of tweets in non-overlapping epochs and dividing by the number of epochs. A more accurate estimate of the average number of tweets at different hours of the day can be acquired using a timeseries model of the activity on the social network.

3.2.2 Timeseries Model of the Dataset

Arrivals entering a system are commonly modelled as a Poisson stochastic process. A Poisson process can be characterized by a random variable $N(t)$ representing the aggregate number of arrivals that has happened up to time t. This random variable has a Poisson

[2] https://dev.twitter.com/docs/streaming-apis/streams/public

probability mass function with a rate λ (as well as two other necessary properties: independent and stationary increment). When the rate is high enough (more than 1000), the Normal distribution can be used as an approximation to the Poisson distribution. Since social media has arrival rates larger than 1000 posts per minute (the time unit is arbitrary but must be long enough to have more then 1000 arrivals), it is therefore possible to assume a normal distribution and use a timeseries model to describe the activity on the social network. One class of timeseries models which is particularly useful for analysis is the class of state space models.

Values of a variable in timeseries data is known to be correlated, and specialized models are developed to deal with such explicit correlation. The explicit autocorrelation (correlation between values of the same variable) is usually uninteresting, since it is known and expected. For example, the data can be known to have a trend such as a chronic increase in the volume velocity. Also, a cyclic increase and decrease in the volume velocity according to the time of day can be expected – this is called the seasonal effect in timeseries analysis. State space methods provide an explicit structural framework for the decomposition of timeseries [22]. Unlike the more popular Box-Jenkins ARIMA models, trend and seasonal effects are not treated as nuisance parameters, and it is not necessary to remove the trend and seasonal effects from the series before the analysis can begin. Instead, a state space model has an equation to capture each type of effect that is suspected to contribute to the values of the observed variable. Different models can be used to describe the same observed variable, and the best model is selected using the Akaike information criterion which penalizes extra parameters.

After experimentation with different models, we found that the model that best describes the volume velocity of Twitter is the following:

1. A deterministic seasonal cycle that is 1 day long, which captures the effect of the hour of day on the number of posts.

2. A stochastic level component, which represents the volume velocity specific to each hour. The stochasticity is represented as a variance attached to the level component.

There was no need to add a trend component, maybe because the sampling used by the Twitter API public stream endpoint emits a stable volume of Tweets, and maybe because there was actually no significant growth in the use of Twitter during the months of data collection. In our model we use the timestamps of tweet creation, and we did not keep the timestamps of when the API returned each tweet as part of the sample.

Figures 3.1 and 3.2 show different components of the model when it is fitted to the hourly number of tweets in the month of October 2012. Figure 3.1 shows the daily cycle component (green) and the noise component (black). The noise component must be present in all state space models, to capture small variabilities in the observed value so that the model does not over-fit the data. The cycle is perfectly aligned with the days in the Hawaii Standard Time (HST) time zone, and thus we always use this timezone (GMT-10) to get the date from a timestamp. Figure 3.2 shows the level component (red) along with a scatter plot of the hourly volume velocity. The blue continuous line is the simple moving average and its confidence bands are the blue dashed lines. The simple moving average is clearly not a a good estimate of the hourly volume velocity. A better estimate of the number of tweets in an upcoming hour can be predicted from the model using a Kalman filter [80].

The peaks in the level component coincide with real world events. The peaks on the 3^{rd}, 16^{th} and 22^{nd} coincide with the presidential debates[3], and the peak on the 28^{th} coincide with the emergency declaration for the northeastern states of the USA in anticipation of Hurricane Sandy[4]. The use of such a model to indicate the presence of interesting information can be investigated further, however it will probably require modelling the number of occurrences of many keywords. This is not scalable because handling correlation between different random variables increases the number of parameters of the model quadratically in the number of variables. The correlation between random variables is captured by a variance-covariance matrix, and if two variables are not known to be independent a parameter must be added in the cell of the intersection of their row and column. We will not use timeseries models any further than this exposition.

[3] http://www.uspresidentialelectionnews.com/2012-debate-schedule/

[4] http://www.fema.gov/hurricane-sandy-timeline#oct28

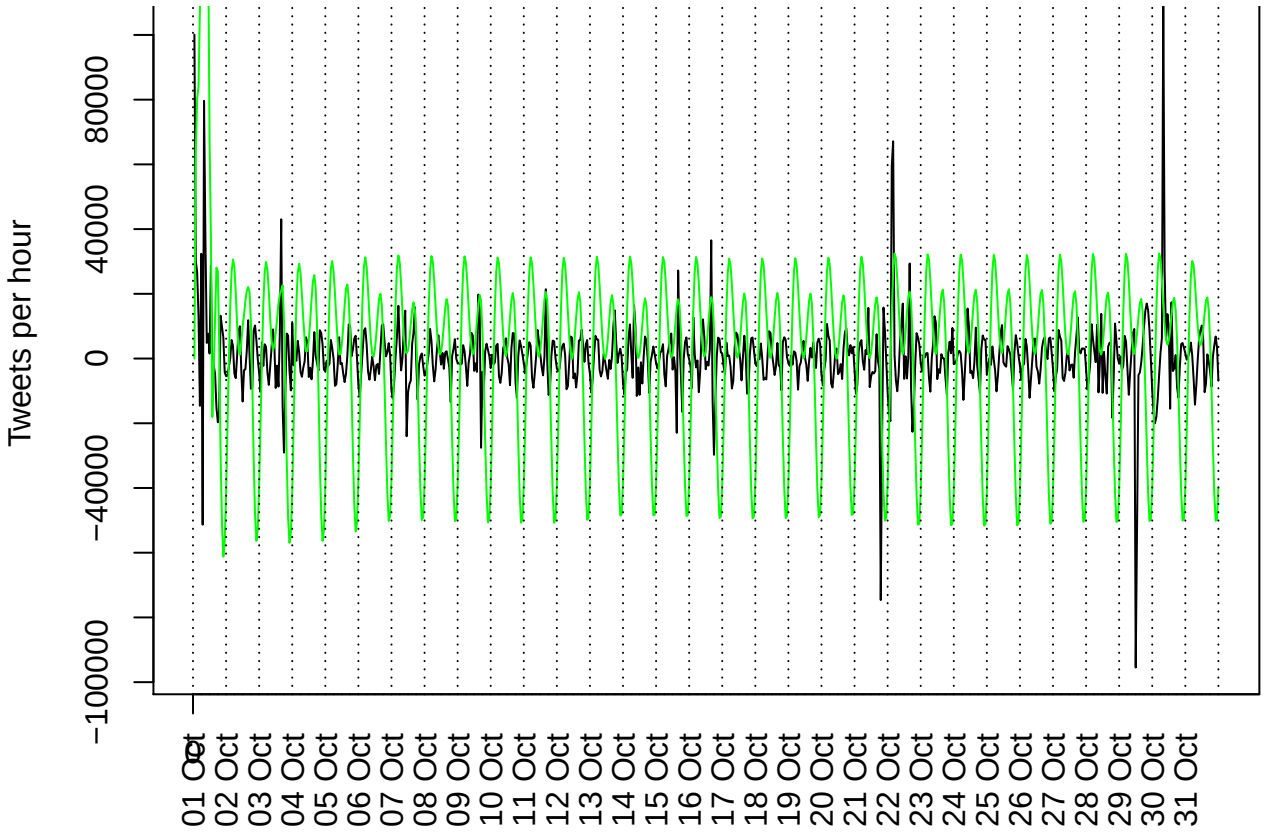


Figure 3.1: Seasonal and noise components of the volume velocity

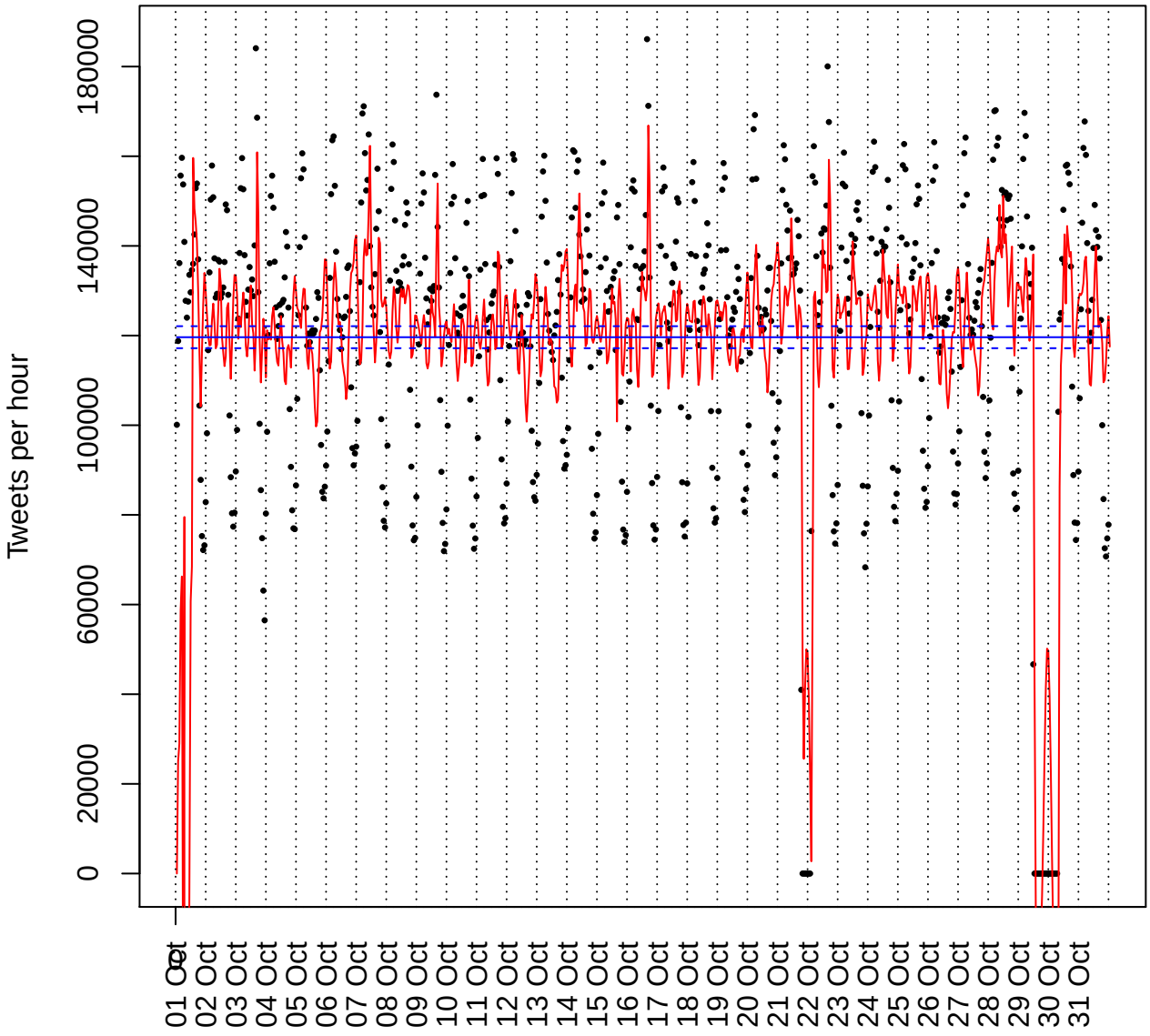


Figure 3.2: Stochastic level of the volume velocity

3.2.3 Dynamic Support Threshold

It is necessary to adapt the support threshold to the dynamicity of the volume velocity at different times of the day. If a ratio support threshold is used then this would not be needed, however determining a support ratio is difficult and it is more desirable to be able to specify the threshold as an absolute number. We therefore define the *minimum support threshold* as the threshold at the hour of the least volume velocity during the day. The *minimum support threshold* is supplied as an absolute number and then converted to a ratio, α, by dividing it by the average of the number of tweets in the hour of minimum volume velocity in different days. The actual support threshold used for mining any given epoch, E, is thus α multiplied by the number of tweets in the epoch:

$$dynamic\_support\_threshold = \alpha \times |E| \qquad where\, \alpha : support\, ratio \tag{3.2}$$

$$= \frac{minimum\_support\_threshold}{\mathbf{avg}(\min_{day}(volume\, velocity))} \times |E| \tag{3.3}$$

Calculating the support threshold dynamically as such makes it intuitive to select a minimum support threshold, but it would lead to a high support threshold in epochs including a burst of tweets. The high support might lead to mining only itemsets about the topic that is causing the burst. Mining a sliding window can alleviate this problem.

3.2.4 Sliding Window Model

One of the prominent stream processing models is the sliding window model, where a window of data *slides* forwards through the stream. The window keeps a fixed amount of history data (fixed in terms of age not size); data with a timestamp older that a certain time is removed from the window as new data is added. The data in the window is processed every time the window slides. Thus, the sliding window model sets a cut-off age after which a datapoint has no effect at all on the results of the stream processing algorithm. This is the simplest way to overcome the problem that an anomaly in the date, such as a burst of tweets about a certain topic, could affect the mining results more than recent data. More elaborate methods include dynamically changing the sliding window size [11], smoothing the data [28], and weighting data according to its age [27].

We apply our algorithms to epochs of data, similar to the *block evolution* stream mining approach [25, 98]. However, our algorithms are not strictly stream processing algorithms. Stream processing algorithms are supposed to process transactions one at a time as they arrive, looking at each transaction only once – in other words, they should make one pass on the data then discard it. This complicates the model of computation and there is no justification to process social media data as a stream. The stream processing model of computation is justified when the data is not supposed to be stored (such as data from sensor networks), or when the processing time has to be as short as possible such that looking at historic data is an overhead (such as in the case of processing stock ticker data, where making a decision a few milliseconds before/after competitors can translate into profits/losses). In the case of social media text, the data has to be stored anyway and the perception of the human users makes processing times in the order of seconds acceptable for a *real time* system.

In essence, processing epochs of data is similar to mining a sliding window that is moved forward by time steps of short span. The time step must be longer than the time needed to mine an epoch of data, and the performance of our algorithms makes it possible to use a time step as short as a few seconds for epochs up to a day long. The window must be moved such that the combinations of transactions not processed together in a batch is minimal. This is achieved by using a time step as short as possible, within the limits of available resources (such as storage space for mining results, if they are retained). Moreover, the time step must be less than the epoch length to avoid missing mining results from a spike happening at the boundary between two epochs.

Figure 3.3 shows the runtime of LCM on epochs of increasing length, and we will show in section 5.1 that our extensions do not degrade its performance. The times reported in figure 3.3 are averages across all epochs of the specified length in the last 3 months of 2012, using a time step that is half the epoch length. The variance is very low and the confidence bands are not shown because they appear as dots. The *minimum support threshold* used throughout this thesis is 10 occurrences in the hour of day in which the volume velocity is the minimum. Figure 3.4 shows the averages of the actual support threshold to which this minimum support threshold translates at different epoch spans.

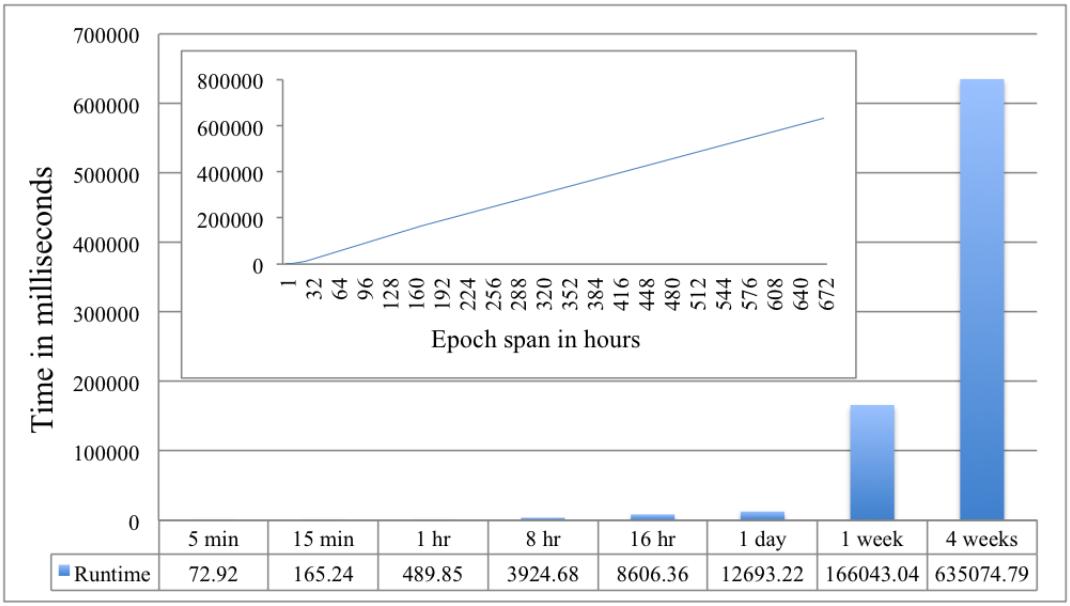


Figure 3.3: Mean runtime of LCM at different epoch spans

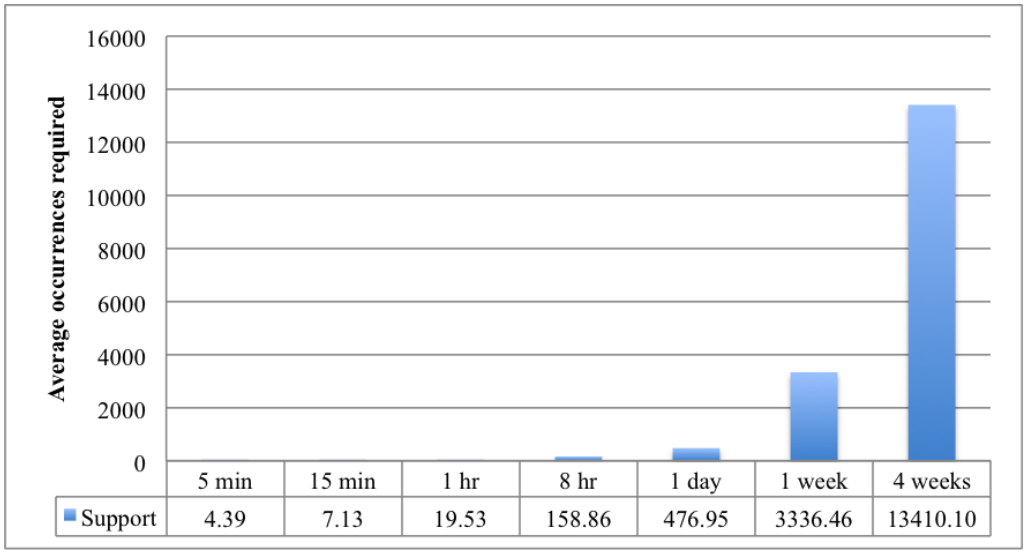


Figure 3.4: Mean support corresponding to minimum support 10, at different epoch spans

In the rest of this paper we mine epochs of 1 hour span. The reason behind this choice is an observation that the number of closed itemsets mined from epochs of span 1 hour or more, at the same minimum support threshold, remains the same. This indicates that itemsets mined from shorter epochs are not included in the results of mining longer epochs. Therefore, the epoch span should be minimized. However, when the epoch span is shorter than an hour the frequency required to surpass the support threshold becomes very low, and the number of mined itemsets increases, with many noise itemsets appearing in the results. For example, the number of itemsets mined from a 15 minutes epoch is double that mined from an hour long epoch.

Regardless of the length of the epoch, many mined itemsets are combinations of function words. In the next section, we outline how we reduce the number of itemsets and eliminate the effect of function words by N-gram filtering.

3.3 Filtering Language Constructs

A large number of itemsets are language constructs that bear no information, such as "such as". By treating sequential language constructs, and any other multiword expression, as one item we eliminate a large number of such itemsets. We can also eliminate itemsets that are made up of all the different fragments of the language construct along with other items; for example, the itemset {we, did, it, #teamobama} can produce 10 other combinations of length 2 or more. There are many measures of association that can be used to detect multiword expressions, but each measure is good only under certain conditions [71]. After preliminary experiments with various measures, we determined that the best performance could be obtained by tokenizing the documents into term N-grams of varying length.

We use term N-gram tokens such that N-grams of high probability are replaced by (N+1)-grams, resulting in a distribution with no high peaks. An N-gram is considered to have a high probability if its maximum likelihood estimate, from a background model, is higher than a threshold η_N. A background language model built from a long epoch of data from the same stream is used for probability estimation. The language model is simply a hash map of the counts of N-grams that appeared within the last month, for N up to 7.

The tokenization of a tweet starts by tokenizing it into unigrams, as described in appendinx A. Then, each unigram of high probability is replaced by two term bigrams, by attaching to it the unigrams before and after it. We keep replacing N-grams of high probability by two (N+1)-grams until there are no more such N-grams.

The threshold of high probability is different for each value of N. The threshold for unigrams is determined as follows: We pick a topical term that is known to steadily appear with a rather high frequency, and is talked about in all languages; for example, 'obama'. The maximum likelihood estimate of the probability of the term 'obama' within the whole collection of tweets is 0.0001. We use $\eta_1 = P(\text{“}obama\text{”}) = 0.0001$. At each N, the probability threshold is adjusted to account for the increase in the number of tokens and the overall increase in the grand sum of counts (caused by overlap). The adjusted η_N is:

$$\eta_N = \eta_1 \times \frac{\sum_{\{w:\, w \in W \text{ and } w.length \leq N\}} freq(w)}{\sum_{\{v:\, v \in W \text{ and } v.length = 1\}} freq(v)} \tag{3.4}$$

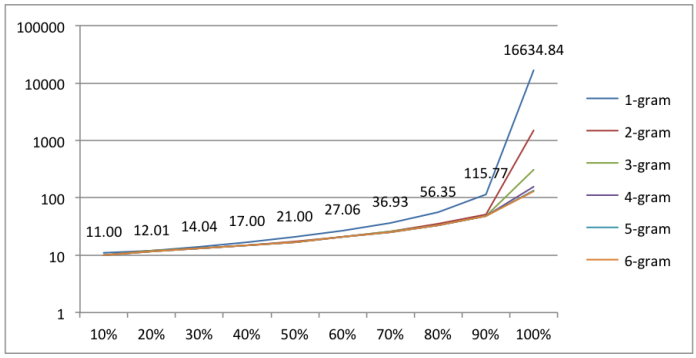
The effect of increasing the maximum length of N-grams is shown in figures 3.5 and 3.6. Figure 3.5 shows the flattening of the distribution by plotting the average decile values at three different times of day. N-grams appearing less than 10 times in an hour (the minimum support threshold we use) are excluded, because they do not contribute to the distribution of items that the algorithm has to mine. The peakedness of the head of the Zipfean distribution is reflected in the peakedness of the 100^{th} percentile. Increasing the maximum N-gram length reduces the peakedness as well as the variance in the peakedness between different hours of day. Figure 3.6 shows how the maximum N-gram length affects the mining algorithm; it shows the number of tokens in the input, the number of closed itemsets, and the runtime of mining one hour of data. The values shown are averages across all one-hour epochs in the month of November 2012. The value of η_1 used is 0.0001. Figure 3.6(a) shows that the number of distinct items increases substantially as the maximum N-gram length increases from 1 to 2, then continues increasing slightly until it starts decreasing at $N \leq 5$. The decrease happens because all 4-grams with probability above the threshold are parts of tweets from services that use the same text and append a URL, such as tweets reporting scores from Game Insight[5]. Such tweets are tokenized into more 4-grams than 5-grams,

[5] http://www.game-insight.com/

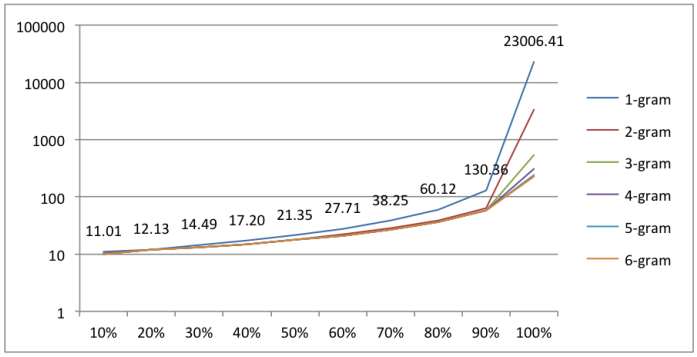
and the 4-grams appearing in them do not appear elsewhere. Thus, some adjacent 4-grams are replaced by fewer 5-grams. Figure 3.6(b) shows that the number of itemsets continues to decrease as expected, with the biggest reduction when the maximum N-gram length increases from 1 to 2. Figure 3.6(c) shows that runtime also decreases as the maximum N-gram length increases from 1 to 5, since LCM's runtime is proportional to the number of closed itemsets, and it can take advantage of the sparsity of data. The runtimes in this figure are slightly less than those in figure 3.3 because they do not include the time taken for writing the posting list of each itemset.

The numbers of itemsets reported in the plots are counts of distinct unigram sets. After mining itemsets of term N-grams we flatten the itemsets to sets of unigrams again. This is necessary since an itemset will have different N-gram set representations, and its postings list is the union of those of the different representations. This also removes overlap between N-grams of the same itemset, making it easier to reason about how itemsets relate to each other. We will use the relation between an itemset and its subsets to select interesting itemsets in the next chapter.

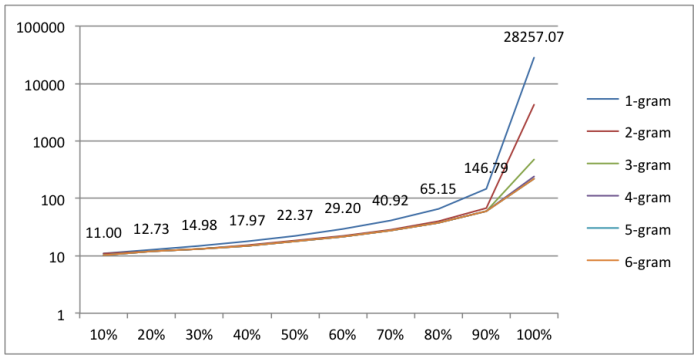
The condition for selecting itemsets that we propose in the next chapter cannot select itemsets of length 1 (frequent unigrams). Thus, to assess its filtering characteristics we have to compare the number of itemsets it selects to the number of itemsets of length 2 or more before its application. At $N \leq 5$, the number of *closed itemsets* of length 2 or more that are mined from an hour long epoch averages at 2,439.17. We also note that the number of *maximal itemsets* of length 2 or more averages at 1,831.92. This high number of *maximal itemsets* show that they cannot substitute the *strongly closed itemsets*, which we propose in the next chapter.



(a) Hour of Day: 00-01 HST (5-6 AM EST)

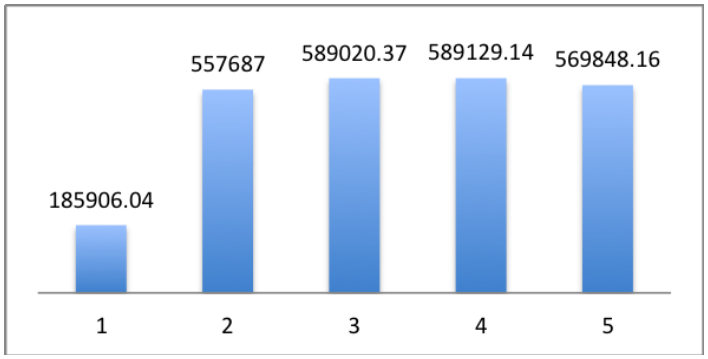


(b) Hour of Day: 08-09 HST (1-2 PM EST)

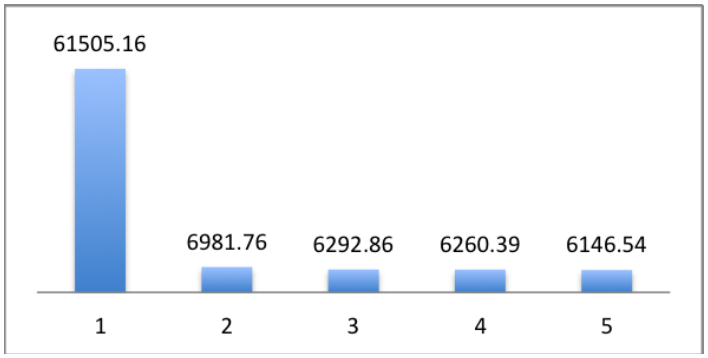


(c) Hour of Day: 16-17 HST (9-10 PM EST)

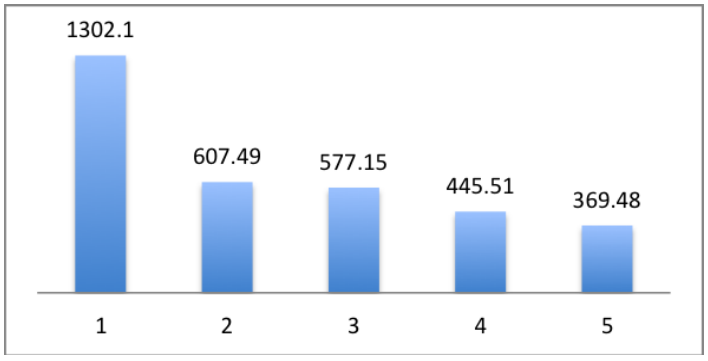
Figure 3.5: Average of token frequency percentiles at different times of day and maximum N-gram length



(a) Mean number of distinct items at different values of maximum N



(b) Mean number of itemsets at different values of maximum N



(c) Mean runtime in milliseconds at different values of maximum N

Figure 3.6: Effect of changing the maximum N-gram length on mining hour long epoches

Chapter 4

Selecting and Ranking Itemsets

In section 3.3, we discussed our handling of function words, using a technique that exploits LCM's tolerance to sparsity. After applying this technique, the average number of closed itemsets mined from an hour of Twitter data drops from 61,505 to 6,146. In this chapter we propose two methods for reducing the number of itemsets even further. The first is a condition for selecting itemsets, and the second is a clustering scheme that merges together similar itemsets to reduce redundancy. We also propose a ranking function that scores itemsets according to their temporal novelty. The two methods and the ranking function all exploit the dynamics of social media, therefore we start by discussing it.

4.1 Social Media Dynamics

As discussed earlier, one use case of social media is to share content that a user finds important with her network. The user may share an original status message about the topic of interest, or *re-share* content posted by other users. In twitter, re-sharing is done by retweeting, which can be done in several ways. One way is by using the "retweet" feature of the Twitter client, which simply forwards the tweet to the user's network. The "retweet" feature keeps a reference to the original tweet in the forwarded tweet, but it does not allow editing it – it automatically prepends "RT: @*orignal_poster_username*". Since

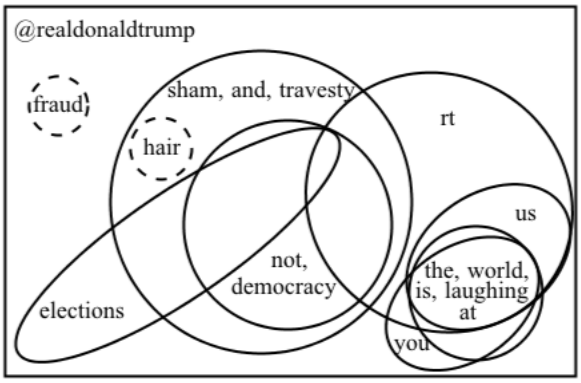


Figure 4.1: Venn diagram of transactions containing items from highly retweeted tweets

editing is not supported by the "retweet" feature, many users chose to retweet manually to be able to add their own thoughts. Recently this can be done using the "reply" feature, which allows a user to write a tweet that references the original tweet thus maintaining the hierarchy of the conversation. However, manually retweeting is still a prevalent way to re-share a tweet along with the user's comment.

There are several conventions for manual retweeting [15]. All the conventions basically quote the original tweet along with its poster's username and a retweet indicator, with the retweeter's comment prepended. Due to the 140 characters limit set on the length of tweets the quotation is usually edited to be as short as possible by selecting only the most discriminative words. This selection is an act of collaborative filtering, but it results in many trivially different subsets from the original tweet, and all subsets with enough support will be mined as itemsets. The additions of retweeters also form many different supersets of the of the original tweet, and some additions may represent opinions that are supported enough to be mined as itemsets.

Figure 4.1 illustrates closed itemsets related to two of Donald Trump's famous tweets in reaction to Obama's victory in the 2012 US elections[1]. Each area in the figure represents the transactions containing the itemset formed by concatenating the items in all intersecting ellipses. The figure shows that one of the tweets was most distinguishable by the words

[1] http://www.huffingtonpost.com/2012/11/07/donald-trump-election-revolution_n_2085864.html

"sham, and, travesty", which are quoted along with Donald Trump's user name in most of the retweets. Other people also chose to include "not, democracy" and/or "elections" in their retweets, and in most of the cases the retweet indicator "rt" was added to the retweet. A few people also added something about Donald Trump's "hair" while retweeting. The second tweet's most discriminative words are "the, world, is, laughing, at". Some people completed this by the original tweet's "us", and others replaced it by "you". At the same time interval, some people tweeted about "fraud" mentioning Donald Trump in their tweets but without any quotation from his tweets.

The figure illustrates how the closed property of an itemset is easily satisfied by an itemset created through the modification of transactions that contain another closed itemset. For example, if a transaction containing a closed itemset is modified by removing one of its items, another closed itemset with enough support is immediately formed resulting in a two closed itemsets instead of one. While an update operation is not supported in the model of frequent itemset mining, a similar effect happens when people retweet, or even when they are writing original posts about a certain fine grained topic. In case of retweeting, people actually start from the original tweet and then remove parts of it to make space for their content. We can imagine that for each fine grained topic there is also a base post that represent the information or ideas within the topic. People make posts about the fine grained topic by selecting parts of the base post and adding their own thoughts. This results in many closed itemsets about the topic that are trivially different from each other.

The fact that any *maximal* itemset is a closed itemset is another *weakness* of the closed condition when used for mining social media text. If a closed itemset is expanded by a certain item a number of times over the support threshold, this results in a maximal itemset which is closed by definition. Now consider that a tweet is retweeted hundreds of times and different groups of people append the same words to it, but no group is considerably large relative to the number of retweets. Since a low support threshold has to be used when mining text, this will result in many maximal itemsets that might not be adding any information to the closed itemset.

4.2 Distinct Itemsets

We have seen that the dynamics of posting on social media result in redundant closed itemsets. To reduce redundancy, we propose the *distinct* condition, which is not as easily satisified as the closed condition, and use it for selecting itemsets. The *distinct* condition builds on the concept of *association rule confidence*. Confidence is the basic property used for association rules mining, and it is used in the definition of δ-free sets [13]. Mining itemsets based on the confidence of rules they induce has long been recognized as a method for finding "interesting patterns" [21], but since this property is not anti-monotone it cannot be directly used to mine itemsets. It has to be used in a post-processing phase. The confidence of an assocation rule that the presence of an itemset, $s_{antecedent}$, implies the presence of another itemset, $s_{consequent}$, is defined as:

$$conf(s_{antecedent} \rightarrow s_{consequent}) = \frac{|T_{s_{antecedent}} \cap T_{s_{consequent}}|}{|T_{s_{antecedent}}|} \tag{4.1}$$

The *distinct* condition is a novel strengthening of the closed condition so that it is not easily satisfied by any modification of a closed itemset. Consider a closed itemset $s_{antecedent}$ that has a particularly high support, as is the case for itemsets made up of the most discriminative words of a certain topic, or a language construct used in different topics. In this case, we can filter out closed supersets of $s_{antecedent}$ that have support high enough to satisfy the support threshold but much lower than $s_{antecedent}$'s support. These itemsets have enough support to be mined as frequent itemsets, but they do not represent a *distinctive* piece of information or opinion within the topic that $s_{antecedent}$ represents.

The *distinct* condition is similar in essence to mining itemsets at multiple minimum supports [54, 74, 84, 91], where a different support threshold is used for each itemset according to the frequency of its components, or its length, or some other function that calculates the *suitable* support threshold. However, *distinct* itemsets are based on a very simple condition that is easy to implement efficiently and intuitive to reason about.

We define a *distinct* itemset as a closed itemset whose frequency comprises more than a certain proportion of the frequency of its least frequent subset – we call this its *parent* itemset. The proportion is a parameter, κ, that controls the selectivity of the condition,

or the *distinctiveness* of itemsets selected by the condition. This can be interpreted as selecting itemsets which are implied by a subset with confidence greater than κ. Formally, the set of *distinct* itemsets, $\mathcal{D}$, is defined as follows:

$$\mathcal{D} = \{s : s \in \mathcal{C}\ and\ \exists\ s_{parent} \subset s\ where\ \frac{|T_s|}{|T_{s_{parent}}|} \geq \kappa\} \tag{4.2}$$

In the mining results of hour long epochs from the Twitter data, the number of *distinct* itemsets, at $\kappa = 0.25$, averages at 347.01 itemsets as compared to 2,439.17 closed itemsets of length 2 or more. To make sure that no important information is filtered out we look at itemsets that are filtered out, and to facilitate the investigation we restrict it to itemsets containing one of the topical terms 'obama' or 'romney' in the first two weeks of November 2012. Excluding epochs in which the terms do not occur at all, the average number of filtered out itemsets that contain a topical term is 7.2 itemsets per epoch. The hour when the US presidential elections result became clear (10 PM EST on 6 November, 2012) happens to be the hour with the most number of such itemsets, because of the large number of posts about 'obama' and 'romney'. Figure 4.2 shows the filtered itemsets from this hour. Most of the itemsets in the figure are not interesting. Furthermore, interesting ones have surrogates in the set of *distinct* itemsets: a subset or a superset or an alternative with similar (sometimes more accurate) information. In the figure, we include below each filtered itemset that we find interesting the *distinct* itemset that we consider its surrogate.

In figure 4.1, *distinct* itemsets are illustrated in solid lines, and closed itemsets that do not satisfy the distinctiveness condition are illustrated in dashed lines. This is not an accurate illustration and it is only meant to give an intuition of how *distinct* itemset are different from closed itemsets which are not *distinct*. We have seen in figure 4.2 some realistic examples of itemsets that are rejected by the *distinct* condition. Figure 4.3 shows the *distinct* itemsets from the same hour at the same κ. Itemsets pertaining to Donald Trump's "sham and travesty" tweet appear on the right side of the figure. Close to the bottom left of the figure there are itemsets showing the number of electoral votes each candidate got at different points of time ({191, 238} and {191, 249}). It is clear from the figure that considerable redundancy remains in *distinct* itemsets. In the next section we propose overcoming this redundancy by clustering similar itemsets together.

1 [romney, that]
2 [still, obama, is]
↪ [still, obama]
3 [romney, what]
4 [romney, with]
5 [will, romney]
6 [romney, if, wins]
7 [romney, for, mitt]
8 [romney, lost]
9 [romney, like]
10 [romney, campaign]
11 [romney, is, mitt]
12 [romney, still]
13 [romney, supporters]
14 [romney, a]
15 [romney, i]
16 [s, romney]
17 [romney, of]
18 [romney, on]
19 [romney, or]
20 [romney, in]
21 [romney, is]
22 [romney, had]
23 [romney, at]
24 [romney, 244]
↪ [obama, 244]
25 [romney, has]
26 [romney, got]
27 [romney, can]
28 [romney, but]
29 [romney, was]
30 [romney, the]
31 [voted, for, obama, who]
32 [2012, obama]
33 [romney, aint]
34 [obama, is, barack]
35 [gana, obama]
36 [obama, gano]
37 [michelle, obama]
39 [years, more, obama, 4]
↪ [#4moreyears, obama]
40 [romney, for, vote]
41 [de, obama]
42 [obama, an]
43 [on, obama]
44 [obama, it]
45 [so, obama]
46 [obama, d]
47 [did, obama, it]
48 [for, obama, it]
49 [que, obama]
50 [URL, obama]
51 [can, obama]
52 [obama, all]
53 [so, obama, won]
↪ [wonnn, obama]
54 [yes, obama]
55 [obama, is, the]
56 [supporters, obama]
57 [obama, when]
58 [with, obama]
59 [obama, team]
60 [obama, has, won]
↪ [obama, we, won]
61 [this, got, obama]
62 [yeah, obama]
63 [romney, about]
64 [obama, happy, won]
↪ [glad, obama, won]
65 [yesss, obama]
66 [president, obama, barack]

Figure 4.2: Itemsets from the hour starting at 10 PM EST on 6 Nov. 2012, which contain 'barack' or 'romeny' and are closed but not *distinct*

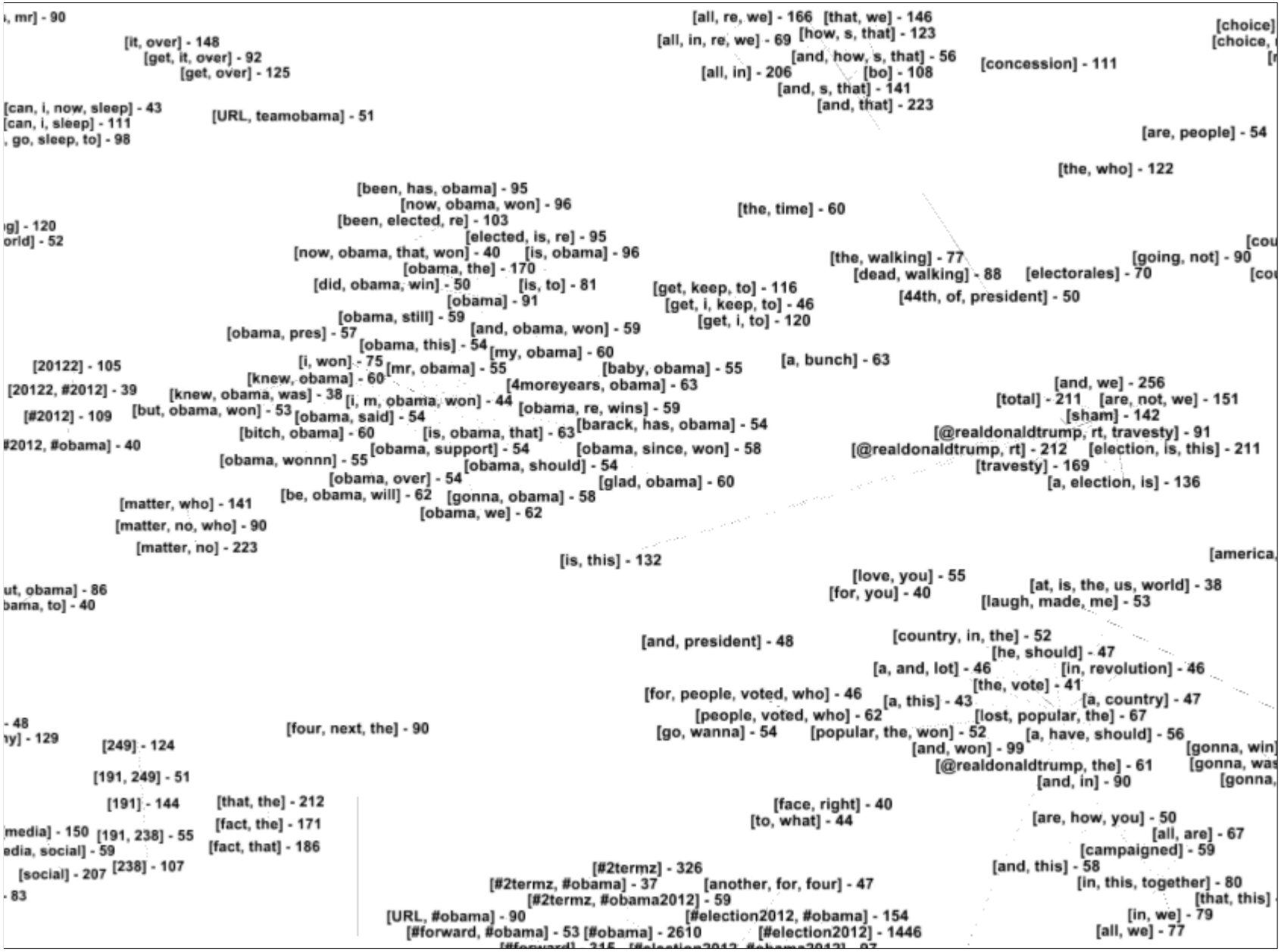


Figure 4.3: Scatter plot of *distinct* itemsets from the hour starting at 10 PM EST on 6 Nov. 2012

4.3 Strongly Closed Itemsets

To overcome the redundancy in *distinct* itemsets we merge similar ones into *strongly closed* itemset clusters. This will also filter out any itemset that does not belong to a particular topic since it will not be part of a cluster. The similarity of a *distinct* itemset, $s_{antecedent}$, and another *distinct* itemset, $s_{consequent}$, is measured as the overlap of the *transactions* containing both of them with the transactions containing $s_{antecedent}$. According to equation 4.1 this is the confidence of the association rule that $s_{antecedent} \rightarrow s_{consequent}$. Confidence is not a symmetric measure, but our proposed clustering makes use of the order in which PPC-extension generates itemsets (as discussed in section 3.1) to determine which itemset should be the antecedent and which should be the consequent.

If PPC-extension uses a total ordering of items that is non-increasing in item frequency, then an itemset generated earlier should be the antecedent, and an itemset generated later should be the consequent. This builds upon the use of variable length N-grams as items, as described in section 3.3, so that the most frequent items are not function words. In that case, following the non-increasing order of item frequency leads to the following:

- Itemsets related to a certain topic are mined close to each other. Notice that PPC-Extension is equivalent to a depth first traversal of a hypothetical prefix tree of closed itemsets. That is, closed supersets of each itemset are generated before moving to an itemset that is not a superset.

- The most discriminative itemset from each topic (according to users' selection) is mined before other itemsets from the same topic, because it is composed of items having the highest support within its topic.

- If an itemset branches out into supersets related to two topics or two opinions within one topic, this itemset is mined before itemsets from either of the two branches. It should not be clustered with either of the branches.

- For an itemset to belong to a topic, it must contain a set of topical words. Since we are processing items in descending order of frequency, then itemsets containing multiple topical words are generated when mining the supersets of the itemset containing only

the most frequent of them. Thus, itemsets belonging to a topic will be mined after the topic's most discriminative itemset, even if they are not its superset.

According to the observations above, the *topical* similarity between itemsets can be indicated by the confidence of the association rule that an itemset generated earlier implies an itemset generated later. Using this similarity measure, we cluster itemsets as follows:

1. Itemsets are processed one by one as they are generated by PPC-extension (using a total ordering that is non-increasing in frequency).
2. If the itemset is not *distinct* it is discarded.
3. A *distinct* itemset, $s_{consequent}$, is clustered with a previously generated itemset, $s_{antecedent}$, if one exists such that the confidence $conf(s_{antecedent} \rightarrow s_{consequent})$ exceeds a threshold, which we take to be $1 - \kappa$ (the indistinctiveness of $s_{consequent}$ from $s_{antecedent}$).
4. If more than one antecedent itemset result in rules with confidence higher than the threshold, the consequent itemset is clustered with the one resulting in the rule with the highest confidence.
5. When an itemset is clustered with another that is already part of a cluster, it is added to the existing cluster.

Each cluster is aggregated into a *strongly closed* itemset, which is the union of all cluster members, and its postings list the union of their postings lists.

This clustering scheme adds an itemset, s_i, to the cluster containing the itemset, s_j, which maximizes the confidence of the rule $conf(s_j \rightarrow s_i)$, with a lower bound on the confidence to maintain distinctiveness. Using $\mathcal{D}$ to denote the set of *distinct* itemsets, we define the desired clustering and the strongly closed itemset represented by each cluster as

follows:

$$\mathcal{R} = \{r : r = \bigcup_i (s_i, s_j) \; where \; s_i \in \mathcal{D} \, and \, s_j \in \mathcal{D}$$
$$and \; \forall_{(s_i,s_j)} \; s_j = \mathbf{argmax}_{s_k} conf(s_k \rightarrow s_i)$$
$$and \; conf(s_j \rightarrow s_i) \geq (1 - \kappa)$$
$$and \; (s_j = r.centroid \; or \; (s_j, r.centroid) \in r)\}$$

$$S_l = \{w : \; w \in \bigcup_{(s_i,s_j) \in r_l} s_i \; where \; r_l \in \mathcal{R}\} \tag{4.3}$$

A different clustering scheme could be used; the goal is to reduce redundancy, and we devised this scheme because it has several merits that makes it suitable. First, the similarity is measured in the transaction space, so that itemsets are clustered according to *topical* similarity and regardless of the lexical similarity. Jaccard similarity is another possible similarity measure, but it cannot be interpreted as the strength of an implication. Also, calculating confidence requires calculating the intersection only, so it is easier to terminate the calculation early if the confidence would not be more than the required threshold. Second, this clustering can be implemented efficiently using techniques similar to the ones proposed by Bayardo et al. [8]. This will be detailed in the next section.

The number of *strongly closed itemsets* from an hour long epoch averages at 139.1 itemsets. This is less than 0.25% of the average number of closed itemsets returned by LCM without any of our proposed methods (61,505.16). *Strongly closed* itemset clusters can be made up of *distinct* itemsets or closed itemsets. The result of the clustering remains the same, but the runtime of the clustering increases by 50% when closed itemsets are used. Besides, when *distinct* itemsets are used it is also possible to include the *distinct* itemsets that are not part of any cluster in the result. This increases the average number of itemsets in the synopsis from 139.1 to 224.49, at $\kappa = 0.25$. *Distinct* itemsets that are not part of any cluster in the Twitter dataset are either language constructs, or named entities that had just enough support to mine them as an itemset, but did not have support that is high enough to make them a centroid of a cluster. We do not rule out the possibility of including them in the synopses, but we chose not to include them in the current work.

4.4 Algorithm for Selecting and Clustering Itemsets

It is straightforward to implement an efficient algorithm for selecting and clustering itemsets as described in the last two sections. The main ideas for an efficient implementation are to limit the comparisons to a few candidates, and to terminate the comparison early if the similarity threshold will not be met. In our case, the postings lists are longer than the itemsets. Thus, we calculate the similarity between the postings lists of two itemsets only if they intersect in one item at least. When calculating the similarity between two postings lists, we can terminate early if the difference exceeds the maximum difference permissible to achieve a similarity of $1-\kappa$, which can be derived from equation 4.1. Algorithm 2 shows a possible implementation.

To illustrate how the algorithm works, in figure 4.4 we show the clustering hierarchy of itemsets pertaining to the "sham and travesty" tweet. The figure shows a tree structure of how itemsets were added to the cluster. An itemset's parent in the tree is the itemset that resulted in the rule with the maximum confidence, which we call the best clustering candidate. An itemset is added to a cluster through its best clustering candidate. The best clustering candidate always has a smaller id, since ids are given sequentially to itemsets and we cluster each itemset with ones that were generated before it. The support of the best candidate is not necessarily higher than the support of the itemset being clustered, but all itemsets within the topic has a support lower than that of the topic's most discriminative itemset, n810={sham} in this case. For example, the hierarchy $n1018 \rightarrow n1019 \rightarrow n1075$ has support values 120, 124, 126 respectively, which are all less than 142 (the support of n810). Finally, we note that this single *strongly closed itemset* cluster replaces 26 *distinct* itemsets. The overall reduction in the number of itemsets will be discussed in section 5.1.

Another example is given in figure 4.5 to illustrate an itemset that branches into different opinions within one topic. These itemsets are mined from tweets of fans voting for different artists to win the MTV Europe Music Awards (MTVEMA). There are 4 distinct clusters in the figure; one for each artist, and one for itemsets not mentioning any artist.

Input: κ: Minimum distinctiveness threshold

Data: $\mathcal{C}$: Closed itemsets produced by LCM, in the order they were generated

Result: $\mathcal{R}$: Strongly closed itemset clusters

for $i \leftarrow 2$ *to* $|\mathcal{C}|$ **do**
  $C \leftarrow \{s_c : s_c \in \mathcal{C}\, and\, c < i\, and\, |s_c \cap s_i| > 0\}$; // Candidates for clustering
  $P \leftarrow \{s_p : s_p \in \mathcal{C}\, and\, p < i\, and\, s_p \cap s_i = s_p\}$; // s_i's subsets (ancestors)
  $s_p \leftarrow \mathbf{argmax}_{s_p \in P} \frac{|T_{s_i}|}{|T_{s_p}|}$; // Direct parent
  if $s_p = null$ ***OR*** $\frac{|T_{s_i}|}{|T_{s_p}|} < \kappa$ **then** // Not a distinct itemset
    $\mathcal{C} \leftarrow \mathcal{C} \setminus s_i$; // Should not be considered in later iterations
    continue ; // Should not be added to a cluster
  end
  $s_m \leftarrow null$, $maxConf \leftarrow 0$; // Best clustering candidate and its score
  foreach $s_c \in C$ **do**
    $\Delta \leftarrow (1 - (1 - \kappa)) \times |T_{s_c}|$; // Maximum difference for confidence 1-κ
    $\delta \leftarrow \text{difference}(T_{s_c}, T_{s_c \cup s_i}, \Delta)$; // Terminates early if $\delta > \Delta$
    if $\delta \leq \Delta$ **then** // The confidence exceeds the threshold
      $conf \leftarrow \frac{|T_{s_c}| - \delta}{|T_{s_c}|}$;
      if $conf > maxConf$ **then**
        $s_m \leftarrow s_k$; // Best clustering candidate
        $maxConf \leftarrow conf$;
      end
    end
  end
  if $s_m \neq null$ **then**
    if $\mathcal{R}[s_m] = null$ **then** $\mathcal{R}[s_m] \leftarrow s_m$; // New cluster with centroid s_m
    $\mathcal{R}[s_i] \leftarrow \mathcal{R}[s_m]$; // Add s_i to the cluster containing s_m
    $\mathcal{R}[s_m].itemset \leftarrow \mathcal{R}[s_m].itemset \cup s_i \cup s_m$;
    $\mathcal{R}[s_m].postingsList \leftarrow \mathcal{R}[s_m].postingsList \cup s_i.postingsList \cup s_m.postingsList$;
  end
end
return $\mathcal{R}$;

Algorithm 2: Forming strongly closed itemset clusters

Id	Itemset	Support
n810	[sham]	142
n906	[and, sham]	135
n907	[and, sham, travesty]	134
n908	[and, sham, total, travesty]	132
n944	[a, and, is, sham, total, travesty]	131
n945	[a, and, election, is, sham, this, total, travesty]	130
n1016	[@realdonaldtrump, this]	129
n1018	[@realdonaldtrump, a, democracy, not, this]	120
n1019	[@realdonaldtrump, a, and, is, sham, this, total, travesty]	124
n1075	[@realdonaldtrump, election, this]	126
n1076	[@realdonaldtrump, a, and, election, is, sham, this, total, travesty]	125
n1055	[a, and, are, democracy, is, not, sham, total, travesty, we]	125
n1056	[a, and, are, democracy, election, is, not, sham, this, total, travesty, we]	124
n1057	[@realdonaldtrump, a, and, are, is, not, sham, this, total, travesty, we]	120
n1079	[@realdonaldtrump, a, and, are, election, is, not, sham, this, total, travesty, we]	121
n1059	[@realdonaldtrump, a, and, are, democracy, is, not, sham, this, total, travesty, we]	119
n1080	[@realdonaldtrump, a, and, are, democracy, election, is, not, sham, this, total, travesty, we]	120
n963	[a, are, not]	132
n964	[a, are, not, we]	128
n1053	[a, and, are, is, not, sham, total, travesty, we]	127
n1054	[a, and, are, election, is, not, sham, this, total, travesty, we]	126
n965	[a, are, democracy, not]	128
n966	[a, are, democracy, not, we]	127
n1033	[travesty, we]	128
n925	[a, is, total]	134
n943	[a, and, sham, travesty]	133

Figure 4.4: The hierarchy of itemsets in the cluster pertaining to the “sham and travesty” tweet

Id	Itemset	Support
n0	#mtvema	624
n2	#mtvema,be,big,the,will	608
n3	#mtvema,be,big,the,will,winner	607
n4	#mtvema,pick,tweet,winner,your	606
n5	#mtvema,at,pick,tweet,winner,your	605
n6	#mtvema,big,tweet,winner,your	605
n11	URL,at,pick,tweet,your	604
n25	#mtvema,be,bieber,big,justin,the,will,winner	405
n26	#mtvema,at,be,bieber,big,justin,pick,the,tweet,will,winner,your	403
n27	#mtvema,be,bieber,justin,think,will	402
n28	#mtvema,URL,at,be,bieber,big,justin,pick,the,tweet,will,winner,your	401
n91	#mtvema,be,big,katy,perry,the,think,will	166
n119	#mtvema,#emawinkaty,be,big,katy,perry,the,think,will	159
n126	#mtvema,#emawinkaty,be,big,i,katy,perry,the,think,will	157
n122	#mtvema,#emawinkaty,be,big,katy,perry,the,think,will,winner	158
n99	#mtvema,be,big,i,katy,perry,the,think,will	164
n104	#mtvema,be,big,i,katy,perry,the,think,will,winner	163
n95	#mtvema,be,big,katy,perry,the,think,will,winner	165
n97	#mtvema,be,big,katy,perry,the,think,tweet,will,winner,your	164
n5207	be,big,gaga,the,will	25
n5275	#mtvema,be,big,gaga,the,tweet,will,winner	24
n5569	#mtvema,at,be,big,gaga,pick,the,tweet,will,winner,your	23
n5562	#mtvema,#emawingaga,at,pick,tweet,winner,your	23
n5572	#mtvema,be,big,gaga,lady,the,tweet,will,winner	23

Figure 4.5: Several clusters for itemsets pertaining to the MTV Europe Music Awards votes

4.4.1 Bounding Space and Time Requirements

Algorithm 2 requires keeping *distinct* itemsets and their postings lists in memory, so that the similarity between a *closed* itemset and all the previously generated *distinct* ones can be efficiently calculated. This is an overhead that the original LCM does not incur, but it is not a large overhead; due to the limited number of *closed* itemsets, the memory requirements does not exceed 6 GB for mining hour long epochs using a Java implementation that uses a lot of objects. However, it is possible to use less memory and also bound the memory requirement by setting a hard limit on the number of itemsets kept in memory.

The number of positions between an itemset and a clustering candidate is 1973.11 on average, with a standard deviation of 1443.64. The large standard deviation indicates that candidates are either found close to the itemset (less than 500 positions behind), or in a position that is several thousands of itemsets behind. According to our observation that itemsets from one topic are generated close to each other, when the total ordering is descending in items frequency, we limit the mining results kept in memory to the last 1000 itemsets and their postings lists. This sets a bound on the memory requirement, as well as the runtime which is dominated by the similarity calculations. Without limiting the number of itemsets kept in memory, the runtime for filtering and clustering hour long epochs at $\kappa = 0.25$ is 5.2 seconds on average. This is reduced to less than 1 second if only the last 1000 results from LCM are kept in memory. More detailed discussion of the runtime performance of the algorithm is given in section 5.1. The *quality* of the *strongly closed itemsets* produced with that limit in place is investigated in section 5.2.

4.5 Temporal Ranking

We now propose a ranking function that scores itemsets according to their novelty, such that the ordered list of itemsets can be presented to a user as a summary of what is happening on Twitter in the mined epoch. The ranks can also be used as weights for itemsets when used by other system components, such as a search component that uses itemsets for query expansion.

We measure temporal novelty compared to a longer period of time leading up to the mined epoch – a background model. A good indicator of novelty is the pointwise Kullback-Leibler Divergence (KLD) between an itemset's probability in the current epoch and in the background model. The KLD of the probability of an itemset, s_i, in the background model, Q, from the current epoch's model, P, can be considered as the Information Gain, IG:

$$\begin{aligned} IG(P(s_i), Q(s_i)) &= -(H(P(s_i)) - H(P(s_i), Q(s_i))) \\ &= \sum P(s_i) \log P(s_i) - \sum P(s_i) \log Q(s_i) \\ &= KLD(P(s_i)||Q(s_i)) \end{aligned} \tag{4.4}$$

This interprets the KLD of an itemset's probability as the number of extra bits in the representation of the itemset used by an encoding based on Q as compared to the number of bits used by an encoding based on P; hence, the information gain in moving from a prior distribution to a posterior distribution.

To calculate the collective IG of itemsets in a *strongly closed itemset*, we have to take into account that the itemsets of the cluster are not independent. For simplicity we will consider only the pairwise dependence between every itemset and the smallest common subset. The joint probability of the smallest common subset, s_{min}, and any itemset in the cluster, s_j, is basically the probability of the itemset s_j, since $s_{min} \subset s_j$. In other words, the probability of s_{min} conditioned on any itemset s_j in the cluster is 1, and there is no information in the occurrence of s_{min} given that s_j occurred. An alternative interpretation is that there is a shared component in the IG of all itemsets in the cluster, which is the IG of s_{min}. Hence, the IG of an itemset cluster, $r = \{s_{i_1}, .., s_{i_m}\}$, can be approximated as the IG of its smallest subset, s_{min}, plus the IG of each itemset in the cluster conditioned on s_{min}:

$$\begin{aligned} IG(P(s_{i_1}, ...s_{i_m}), Q(s_{i_1}, ...s_{i_m})) &= IG(P(s_{min}), Q(s_{min})) \\ &+ \sum_{j=i_1..i_m} IG(P(s_j|s_{min}), Q(s_j|s_{min})) \end{aligned} \tag{4.5}$$

This formula avoids implicitly adding the information gain from s_{min} multiple times to the total information gain. It can be used for ranking clusters, but it will favour larger clusters.

We normalize by the size of the cluster, giving our final ranking formula for strongly closed itemsets:

$$score(r) = IG(P(s_{min}), Q(s_{min})) + \frac{\sum_{j=i_1..i_m} IG(P(s_j|s_{min}), Q(s_j|s_{min}))}{m} \tag{4.6}$$

In this thesis, the background model we use for each day is the results of mining the 4 weeks before it, using a *minimum support* value of 1 occurrence. Mining the background model at such a low support increases the number of produced itemsets, which is desirable for a background model. All probability estimates are smoothed by add-one smoothing.

Chapter 5

Evaluation

In this chapter we analyze the performance of the methods proposed by applying them to the mining results of all one-hour long epochs in the Twitter dataset. We evaluate their efficiency in reducing the number of mined itemsets in section 5.1, and we assess their effectiveness for choosing a high quality synopsis in section 5.2. The experiments were run using a single threaded Java implementation of algorithm 2, running on a 1.4 GHz processor with 2 MB of cache.

5.1 Efficiency

We evaluate the efficiency of our proposed filtering and clustering methods in terms of the number of itemsets before and after their application, and the runtime overhead they add on top of the runtime of the LCM algorithm. Using variable length N-grams as items, the average number of itemsets mined from an hour-long epoch is 2,439.17 closed itemsets of length 2 or more after flattening; that is, excluding itemsets that are merely a frequent unigram. We start by analyzing the reduction in this number at different values of the distinctiveness threshold, κ.

Figure 5.1 show the effect of varying κ on the mean number of *distinct* and *strongly closed* itemsets. The number of *distinct* itemsets drops as the distinctiveness threshold

increases. On the other hand, the number of *strongly closed* clusters formed increases as the similarity/indistinctiveness threshold (taken as $1-\kappa$) decreases. The dashed line shows that the number of unclustered distinct itemsets reaches zero at $\kappa = 0.5$, explaining why the number of clusters changes very slightly after that. We use $\kappa = 0.25$ in this thesis, which is an arbitrary choice based on the definition not the data. The average number of *strongly closed* itemsets mined from one-hour epochs at this value of κ is 139.1, which is about 5.7% of the number of closed itemsets of length 2 or more. If we also include unclustered *distinct* itemsets in the synopses, their average size increases to 224.49 itemsets per hour long epoch.

Figure 5.2 shows the total runtime of the LCM algorithm plus filtering based on the distinct condition and clustering into strong closed itemsets at different epoch spans. The runtime of LCM alone is also plotted for reference. We also plot the performance of another frequent itemset mining algorithm, FP-Zhu [29], which was the runner up at FIMI 2004 [41]. We include it to show that our extensions do not degrade the performance of LCM even in the context of competitions. The y-Axis is in logarithmic scale to keep the scale of the plot suitable for seeing slight differences. The output of LCM is the input to the filtering and clustering step, so it is affected by the number of closed itemsets produced. Recall that the number of itemsets from epochs of span less than an hour is higher than that from epoch of span one hour or more. This explains why it takes slightly longer time for clustering results from the 15-minute epoch and then takes a constant time for epochs of a longer span.

5.2 Effectiveness

We now show examples of the synopses produced by ranking the *strongly closed* itemsets as discussed in section 4.5. The examples will serve as our assessment of the quality of the synopses. We make an effort to choose epochs from time periods when it is known what people should be talking about. In section 5.2.1, we show how the 2012 US presidential elections day is summarized by our methods. We also compare our summary to the one produced by a state of the art algorithm, which generates better quality summaries than

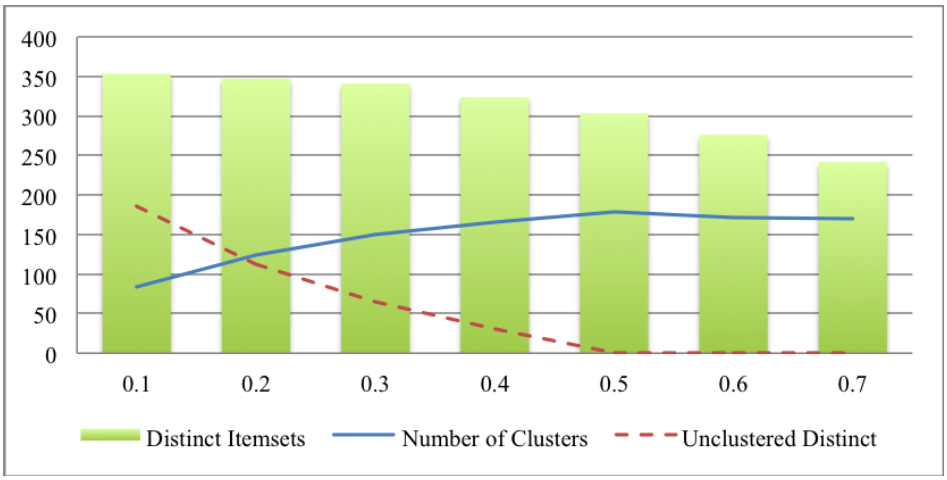


Figure 5.1: The number of different types of itemsets at different values of κ, the distinctiveness threshold. The value of κ used in this thesis is 0.25, midway between 0 and 0.5 when all distinct itemsets are clustered, but it was determined according to the definition.

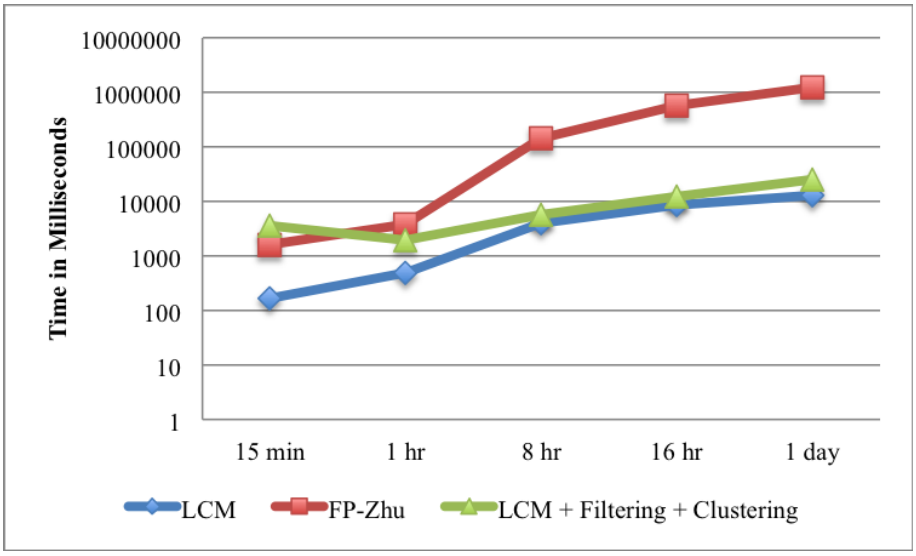


Figure 5.2: Runtime, at different epoch spans, of the LCM algorithm with and without the filtering and clustering extensions as well as of another efficient frequent itemset mining algorithm. Our extensions do not degrade the efficiency of the LCM algorithm.

other summarization algorithms based on itemsets [57]. In section 5.2.2, we investigate the synopses of hours selected by the aid of Google Trends[1], assessing which popular queries were or were not reflected in our synopses.

5.2.1 U.S. Presidential Elections Day

In this section, we focus on the 2012 US presidential elections day, November 6^{th}. We compare the quality of the synopsis of the day created by our proposed methods to the quality of the synopsis created by mining the top itemsets with the MTV algorithm [57] (according to a maximum entropy model). The synopsis should be suitable for giving a user an overview of what people were Tweeting about at different times of the day.

Table 5.1 shows the top 3 *strongly closed* itemsets from one-hour epochs on the elections day, using a half-hour time step. The itemsets shown are the first appearances of the most interesting itemsets; that is, an hour is shown only if its top 3 feature novel interesting itemsets[2]. The first column is the beginning of the hour, EST time. The second column is the top itemsets, reordered and punctuated to be meaningful phrases by looking at a few tweets from the postings list of each itemset. The third column is a commentary to explain the itemsets, also composed constructed according to tweets in the postings lists. The fourth column is the rank of an equivalent itemset in the top 30 itemsets mined by the MTV algorithm, if one exists.

We can see how the events of the US presidential elections unwind from "get out and vote" to the projections and debates, all the way to the "acceptance speech". Early in the day, people were excited and congratulating each other for being on the elections day. The excitement continues for one more hour, with some people taking it too far by sending pictures of their ballots. Itemsets about the elections stop occupying the top 3 ranks until the evening. During the day, the top 3 ranks were occupied by itemsets about UEFA Champions football matches (with timely updates of their scores), and about TV shows.

[1] http://www.google.ca/trends

[2] The top 50 itemsets in the hours from 5 AM on the 6th till almost 9 AM on the 7th are available for download at http://plg.uwaterloo.ca/~yaboulna/thesis_results/twitter_synopses/elections-day_nov6-0500_nov7-0850_top50-strongly-closed.txt

We include the top 3 itemsets of 3 PM as an example. In the evening, the elections heat up and events about it keep occupying top positions in all epochs. At 10 PM all the top 5 were relevant and very interesting that we decided to include an extra two itemsets in the example (we do this for MTV as well, in the hour of the acceptance speech). Shortly after the results of the elections became clear, news that Marijuana and same sex marriage were legalized in some states occupy the top rank. This exemplifies the power of social media as a collaborative filter, selecting the news of greatest importance to social media users. The user centric definition of importance is also evident in the attention given to the "lady with a flag in her hair" appearing on TV behind Obama during the acceptance speech.

The top 3 itemset mined by the MTV algorithm [57] are usually itemsets from Tweets sent by services reporting users' scores in games, or the number of new followers each user got. However, the top 10 itemsets usually contain interesting itemsets[3]. Therefore, in table 5.2 we show itemsets that we find interesting from MTV's top 10, assuming that the itemsets from Tweets sent by automatic services can be filtered out by another dedicated filter. The actual rank in which the itemset appeared is shown in column 3, and the rank in which it appears among *strongly closed* itemsets is shown in column 4. We show only hours in which the top 10 itemsets contain novel interesting itemsets. The itemsets chosen by MTV tend to be complete Tweets which were retweeted by many users, thus we omit the explanation column for brevity. The input to the algorithm was transactions made up of N-grams up to 5 terms long. This helped the algorithm converge faster since the distribution is flatter. The use of N-grams also overcomes the dominance of language constructs, which are ranked high in all hours if unigrams are used.

All the topically relevant itemsets chosen by MTV are present in the top 50 *strongly closed* itemsets – most of them are in the top 30. We do not know if all the top ranked *strongly closed* itemsets would also be picked by MTV if we use it to mine more itemsets. The space and time requirements of MTV becomes prohibitively large when the number of itemsets requested is increased, as it takes more than an hour to mine the top 50 itemset from an hour long epoch. At $k > 100$ the MTV algorithm fails because of a memory error,

[3]The top 30 itemsets in the hours from 5 AM on the 6th till 5 AM on the 7th are available for download at http://plg.uwaterloo.ca/~yaboulna/thesis_results/twitter_synopses/elections-day_nov6-0500_nov7-0500_top30-MTV.txt

Time	Itemset	Explanation	MTV
08:30	*get out and vote*	Encouraging participation	17
	happy elections day	Congratulations	27
	it's elections day	Excitement	-
09:30	*if romney wins*	People reflecting about the different possible results	-
	of your ballot	Telling people to stop sending pictures of ballots	-
	in line to vote	Tweeting while standing in line	28
15:00	*Borussia Dortmund Real Madrid 2*	Score update of UEFA football match	17
	incroyable talent	France's version of the Got Talent TV show	-
	City Ajax 2	Score of Manchester City vs Ajax	-
19:30	*Linda McMahon*	Linda McMahon loses CT senate race	-
	Florida is so ...	tight or close; the number of votes	-
	Romney is winning	Speculations about the results	-
20:00	*New Hampshire*	Obama won in NH	9
	Obama to win / is winning	Two itemsets speculating	-
	too close to call	The results are still not clear	22
21:00	*Ohio and Florida*	Obama won in 2 more states	-
	Sarah Palin	Sarah Palin appears on Fox news	-
	concession speech	Todd Akin's concession speech	-
22:00	*The electoral college*	Republicans complaining about the voting system	9
	President of the United States	Barack Obama is the 44th president of the USA	8
	who voted for	People turning against each other	-
	once you go black you never go [back]	A popular culture reference	15
	my president is black	Americans proud about the equality of opportunities	-
22:30	*@realdonaldtrump this elections (see table 5.2 for the rest of the tweet)*	One of Donald Trump's famous melt down tweets. This cluster merges 26 *distinct* itemsets.	1
	concede to	The losing candidate has to concede	-
	Obama won. I ...	People's reactions to the results	-
23:00	*same sex marriage*	Some states legalized same sex marriage	27
	for recreational use	Colorado legalized Marijuana for recreational use	21
	acceptance speech	People anticipating the speech	26
00:30	*Delivered, sealed, signed*	Stevie Wonder's song used in Obama's campaigns	10
	The best is yet to come	Quote from Obama's acceptance speech	2
	with the flag in her [hair]	During the speech, attention on social media goes to a woman appearing behind Obama on TV!	14
01:00	*Four more years*	Barack Obama's tweet; the most retweeted in 2012	20
	@paulmyers47: spots	Pyramidal marketing; people retweet to make money	12
	President Barack Obama	News headlines about the speech uses the formal title	-

Table 5.1: Top 3 *strongly closed* itemsets for hours in US presidential elections day

Time	Itemset	Ranks	
19:00	*A partir de que idade você considera alguém velho? (Portuguese for "From what age do you consider someone old?")*	1	22
	A qué edad consideras que alguien es viejo? (Spanish for the same question above; we do not know what started this meme)	4	10
	We'll find out if it's true. America went black, will it go back?	9	24
20:30	*Obama rhymes with Ohana. Ohana means family & family means nobody gets left behind. Mitt rhymes with shit.*	3	3
	New Hampshire	9	1
	#IamSickOf	10	-
21:00	*We're all in this together. That's how we campaigned, and that's who we are. Thank you. -bo*	1	22
	@barackobama This happened because of you, thank you.	4	20
	@barackobama Four more years URL	6	6
22:00	*@realDonaldTrump: We can't let this happen. We should march on Washington and stop this travesty. Our nation is totally divided!*	5	44
	President of the United States	8	2
	@realDonaldTrump: The electoral college is a disaster for a democracy.	9	1
22:30	*@realDonaldTrump: This election is a total sham and a travesty. We are not a democracy!*	1	1
	@realDonaldTrump: Our country is now in serious and unprecedented trouble...like never before.	3	24
	@realDonaldTrump: He lost the popular vote by a lot and won the election. We should have a revolution in this country!	5	31
23:00	*@realDonaldTrump: Lets fight like hell and stop this great & disgusting injustice! The world is laughing at us.*	4	48
	@realDonaldTrump: House of Representatives shouldn't give anything to Obama unless he terminates Obamacare.	6	15
	*@adamlevine: That's what happens when you fu** with Sesame Street*	8	18
00:30	*The best is yet to come*	2	2
	Michelle, I have never loved you more	5	12
	We are an American Family, and we rise ...	8	20
	Delivered, sealed, signed, I'm yours	10	1

Table 5.2: The most interesting itemsets out of the top 10 itemsets picked by the MTV algorithm for hours in the US presidential elections day. Column 3 is the actual rank given to each itemset by MTV. Column 4 is the rank of an equivalent *strongly closed* itemset.

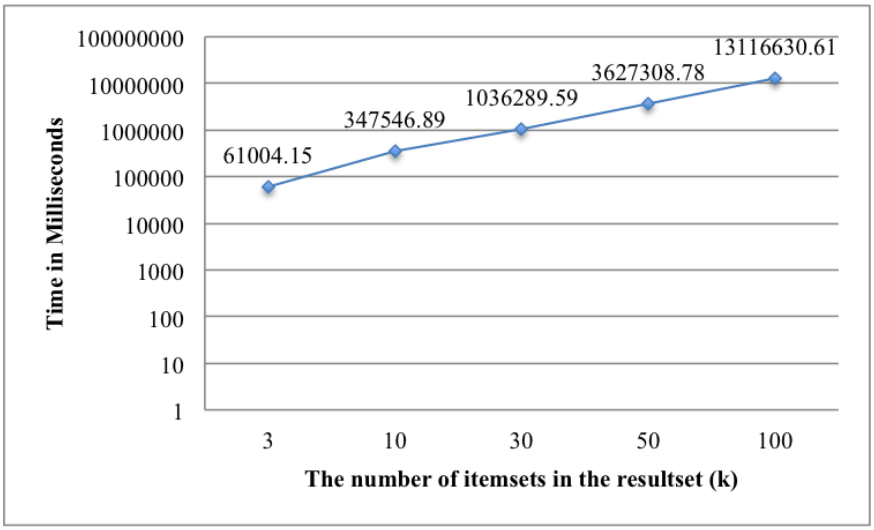


Figure 5.3: Runtime of mining an hour long epoch using the MTV algorithm [57] at different resultset sizes (top k). The runtime increases exponentially with increasing values of k.

even though it was being run on a machine with 256 GB of RAM, with no limit on its resource consumption. The runtime performance at $k \leq 100$ is plotted in figure 5.3. The small value of k that has to be used makes MTV inadequate for mining social media, where the high volume of posts and the variety of topics users talk about is likely to result in more itemsets than what MTV would pick as the top k. On the other hand, *strongly closed* itemsets are efficiently mined and are not restricted to a predefined number.

5.2.2 Google Trends

Google Trends is a freely available service provided by Google. It analyzes queries submitted to the Google search engine by millions of users worldwide, and it can be used to know the most popular queries for different geographies[4]. It can also be used to compute trends of user specified queries over time[5]. Queries that are named entities according to

[4] http://www.google.ca/trends/

[5] https://support.google.com/trends/answer/87276?hl=en

Google's Knowledge Graph[6] are automatically tracked and aggregated by category into Top Charts[7] – lists of real-world people, places and things ranked in order of search volume during user specified intervals of time[8]. We use the Top Charts of November 2012 for the people category[9] (only available for queries made from the USA) as an aid for finding out what social media users are expected to be talking about. We observed that named entities that make it to the top 10 chart in the people category are divided into finer groups. The queries about people in each group have similar volumes over time, with peaks at the times of real world events in which those people in the group took part. For example, the top two most searched for people are Barack Obama and Mitt Romney. Figure 5.2.2 is the plot showing the volume of queries about either of them, as displayed by Google trends[10]. The two curves in the plot are similar; both have peaks at the elections day and are otherwise flat. Note that the y-Axis in the figure is scaled[11] so that the maximum volume per day for either of the queries becomes 100 (to our surprise the curve for "Barack Obama" is much lower that that of "Mitt Romney", even though they occupy positions 1 and 2 respectively in the Top Chart). Since entities can be grouped together according to the events which made users search for them, we use the Top Charts to know about real world events which were of interest to internet users. We evaluate the use of *strongly closed* itemsets as temporal synopses by investigating the mining results from the time periods when these events were happening. We look for itemsets about those entities and events in the top 30 *strongly closed* itemset, and assess which of them were or were not reflected in our synopses. The number of itemsets from each hour is limited to 30 because this is a number that can be conveniently displayed as a summary for a human user.

[6] http://www.google.com/insidesearch/features/search/knowledge.html

[7] http://www.google.ca/trends/topcharts

[8] https://support.google.com/trends/answer/3076011?hl=en

[9] http://www.google.ca/trends/topcharts#vm=chart&cid=people&geo=US&date=201211

[10] http://www.google.ca/trends/explore#q=Barack%20Obama%2C%20Mitt%20Romney&geo=US&date=11%2F2012%201m&cmpt=q

[11] https://support.google.com/trends/answer/87282?hl=en&ref_topic=13975

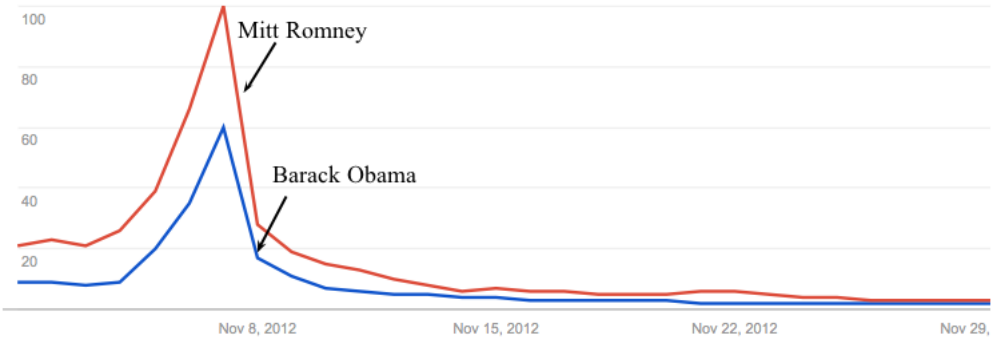


Figure 5.4: Google Trends' volume with time plot for the queries "Barack Obama" and "Mitt Romney", the top 2 in the Top Chart for the *people* category in November 2012. Both curves show a peak of interest at the day of the U.S. presidential elections.

Pop artists and music awards events

Pop artists occupy most of the top 10 chart for the *people* category. Five out of the top 10 are pop artists, and they occupy ranks as high as 3 (Justin Bieber) and 4 (Taylor Swift). This is due to music awards events, specially the MTV Europe Music Awards on November 11th[12] and the American Music Awards on November 18th[13]. Figure 5.2.2 shows the query volumes for the artists in the top 10 chart. The curves for all of the artists show relatively high volume when those events were happening.

The top 30 *strongly closed* itemsets mined from hours leading to the MTV Europe Music Awards contain many itemsets about the events and the artists, especially that this event used social media to collect audience votes. Voting for the award winner is a good example of a topic where people have strongly different opinions. We have shown earlier, in figure 4.5, that such different opinions are all reported as separate *strongly closed* itemsets, because clustering in the transaction space avoids forming incohesive clusters. We also note that different *strongly closed* itemsets from one such topic can all occupy high ranks. For example, the top 10 itemsets for the hour starting 9:30 EST on November 11^{th}

[12] http://en.wikipedia.org/wiki/2012_MTV_Europe_Music_Awards
[13] http://en.wikipedia.org/wiki/American_Music_Awards_of_2012

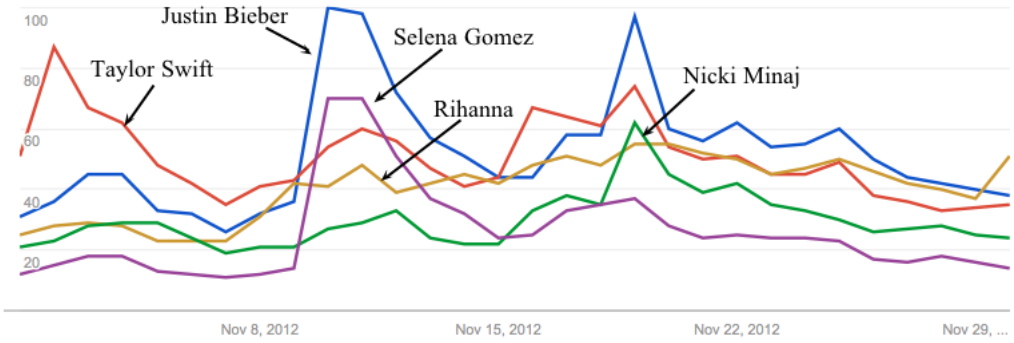


Figure 5.5: Google Trends' volume with time plot for the queries about pop artists in the Top Chart for the *people* category in November 2012. The curves for different artists show similar trends with high search frequency at the times of music awards events.

are shown in table 5.2.2. The number in parenthesis next to each item is the number of itemsets in the cluster which contain this item. Items in the itemset are sorted by their frequency in this hour, in descending order. The first itemset in the top 10 reflect the winner, Justin Bieber. Itemsets about other topics also appear in the top 10; there was a football game between Chelsea and Liverpool, and November 11^{th} is the day World War I ended (observed as Remembrance day in the United Kingdom and the Commonwealth of Nations, and Veteran's day in the USA). Several other hours have itemsets supporting different artists in the top 30, which are all variations of the itemset: {i, think, ARTIST, NAME, will, be, the, big, winner, tweet, your, pick, at, URL[14], #mtvema}.

The American Music Awards (AMAs) did not use social media to collect votes, so there was not as much posts about it as MTVEMA. However, itemsets about it still occupied top positions. There are several hours where the itemset {american, music, awards} occupied the top rank, but the hour we choose to show as an example does not. We chose this hour because it contains an itemset that offers more information than what is contained in other itemsets from other hours; itemsets during this event were mainly names of artists, or the name of the event. The itemsets from this hour, shown in table 5.2.2, contains an itemset (at position 26) about how Justin Bieber appeared with his mother Pattie

[14] http://ema-twittertracker.mtv.com/live/predict.html

1	#emawinbieber(43), #mtvema(92), URL(31), at(47), be(77), bieber(62), big(81), i(24), justin(62), pick(50), the(62), think(48), tweet(74), will(77), winner(67), your(65)
2	#jonasbrothers(3), #emawinjoej(4), #joejonas(3)
3	#emawinbieber(1), #justinbieber(2)
4	#emawinbieber(2), #justinbieber(2), #emavotebieber(4)
5	chelsea(2), liverpool(2), vs(3)
6	#staycurrent(3), you(1), @buysocialclout(2), can(1), follow(1), today(1)
7	#emavoteonedirection(1), URL(2), emavoteonedirection(1)
8	#emawinbieber(1), #justinbieber(2), @justinbieber(2)
9	remembrance(2), day(1)
10	veteran(2), day(1), s(2)

Table 5.3: Top 10 itemsets in an hour in the morning of the day of MTV Europe Music Awards, showing different Strongly Closed itemset clusters voting for different artists

Mallette[15] instead of his usual escort Selena Gomez, his ex-girlfriend. This is interesting news, specially following Justin and Selena's break up that warranted both of their names a surge in their popularity as search queries on November 9^{th} and 10^{th}. The news of the break up appeared in the top 30 itemsets of different hours on those days. We include some example tweets randomly selected from the tweets in the postings lists of the *strongly closed* itemset {justin, bieber, and, selena, gomez}. This shows how including the postings lists in the mining results makes the synopses particularly rich, because the actual posts from which they were mined can be easily retrieved:

- *"Justin Bieber and Selena Gomez Break Up http://eonli.ne/SByVNK" WHAT?!?!?! Is this truee? @justinbieber @selenagomez*"
- *"Justin and Selena broke up? O_o the awkward moment when he wrote a song about her and another with her name in it..."*
- *"apparently there is a rumor that justin bieber and selena gomez broke up? GO @jessleigh1122 GO!!!!!"*.

[15] http://hollywoodlife.com/2012/11/18/justin-bieber-brings-mom-to-amas-without-selena-gomez/#!1/nokia-jenny-mccarthy/

The itemsets in table 5.2.2 also include the names of other artists that are not in the Top Chart. Liam Payne (rank number 5) is a member of the band One Direction, which won several awards that night. Carly Rae Jepsen (rank number 9) also performed and won an award at the American Music Awards[16].

People appearing in the Top Charts but not in the synopses

Out of the five pop artists appearing in the Top Chart, as shown in figure 5.2.2, the names of only four appeared repeatedly in the top 30 itemsets of hours in the days of their respective higher search volume. The name of Rihanna never appeared in any *strongly closed* itemset. Actually it rarely appeared in any itemset of length 2 or more. It is not unusual for an itemset about an artist to be merely his or her name, and several artists, like Rihanna, are known by their first name only. It is impossible for an itemset of length 1 to be *distinct*, since this condition requires that the itemsets be implied by one of its subsets with high confidence. Therefore *distinct* and *strongly closed* itemsets would never capture information in an itemset of length 1. The same happens with one more of the names in the Top Chart, Abraham Lincoln. During the month of November people were talking on Twitter about Lincoln the movie and this was reflected in mining the term as a closed 1-itemset from several hours throughout the month. However, there were not any closed itemsets longer than 1 item, and thus there was no mention of Lincoln the movie in the synopses. This shortcoming can be overcome by settitng a lower support threshold so that longer itemsets are mined.

Another two celebrities who appear in the top 10 chart but not the synopses are Kim Kardashian and Michael Jordan. The volume of queries about Kim Kardashian is different from other artists, since she is not a pop music artist and does not participate in music awards events. For these two celebrities, the volume of queries does not have high peaks, and the curve is relatively flat. Itemsets mined from Twitter rarely mention either of them. Of all itemsets mined from hour long epochs throughout the month of November, only 37 closed itemsets contain “Kim Kardashian” or “@kimkardashian”, and none contain

[16]http://blog.muchmusic.com/carly-rae-jepsen-wins-at-the-american-music-awards-puts-on-fun-and-flirty-performance/

1	@vinoalan(1), #vinofollowspree(2)
2	vote(2), #peopleschoice(2), voted(1), retweet(2), @peopleschoice(4), URL(3), for(1), i(1), just(1), to(2), via(3)
3	@simoncowell(2), follow(2), me(2), please(2), to(1)
4	grinch(2), stole(2), all(3), christmas(2), for(2), haters(3), how(1), is(1), the(5), this(1)
5	you(1), @real_liam_payne(2), thank(2)
6	grinch(3), stole(2), how(1), the(3)
7	all(3), amas(1), for(3), is(2), the(4), this(2), watching(2)
8	grinch(1), stole(2), christmas(2)
9	jepsen(3), carly(1), rae(2)
10	dumb(1), dumber(2)
11	3(2), @real_liam_payne(2), URL(1), live(1), on(1)
12	hi(2), say(1), to(2)
13	oz(1), wizard(2), of(2), the(2)
14	liam(1), @real_liam_payne(1), URL(1), ama(1), live(1), on(2), the(1)
15	american(2), awards(2), music(3)
16	ama(3), award(3), first(1), night(1), of(3), s(2), the(6), watching(1)
17	@simoncowell(1), @real_liam_payne(1), call(2), follow(2), me(3), so(2)
18	pone(3), triste(2), qu(1), te(3)
19	is(1), on(1), red(1), the(2), wanted(1)
20	@simoncowell(1), simon(3), call(1), follow(3), me(2), please(3)
21	american(2), awards(1), music(2), the(2)
22	carpet(2), red(1)
23	oz(1), wizard(3), of(2)
24	imessage(2), is(1)
25	series(1), survivor(2)
26	and(3), justin(2), pattie(2)
27	you(4), @real_liam_payne(3), URL(1), found(2), i(2), live(1), love(1), on(1)
28	#getglue(2), URL(1)
29	#gameinsight(1), #android(2), #androidgames(1), URL(3), android(1)
30	#gameinsight(1), #android(2), #androidgames(3)

Table 5.4: Top 30 itemsets from 18:30 EST on the day of American Music Awards (AMAs) showing several mentions of the event, as well as news about Justin Bieber in rank 26.

"Michael Jordan" or "@jumpman23" (his verified username). This means that the number of occurrences of their names is lower than the support threshold in the all or almost all of the 1,440 overlapping hour epochs in the month. We do not know why is it the case that the interest shown by Google search users was not reflected by Twitter users, or maybe this is an artifact of the algorithms used to enumerate the Top Charts.

Chapter 6

Conclusion and Future Work

In this thesis we have proposed a method for efficiently creating temporal synposes of social media streams, based on a frequent itemset mining algorithm that is suitable for sparse data, LCM. The synposes can be used to explore social media, and they can also reflect what is going on in the real world. Several studies have shown that social media is a good source of information about real world events, offering timely information from a variety of sources. Actually some of the information on social media is not available in traditional news media, or might be available on social media first then get picked up by news media. Social media allows citizen journalists and ordinary people to make posts that can get disseminated widely based on the quality of the post, and regardless of the fame of the poster. Our methods focus of the content of posts, without explicitly giving different weights to posts made by different types of account owners. Thus, the synopses reflect the interests of social media users at different points of time, giving voice to ordinary people and allowing the dynamics of social media to surface content which users find intriguing.

6.1 Contributions

Our method summarizes an hour of Twitter data (119,035.49 tweets on average) into 139.1 itemsets in 1,945.68 milliseconds on average, and scales well for longer epochs of data. The

direct application of LCM on one-hour epochs of Twitter data results in an average of 61,505.16 closed itemsets and takes 2,506.58 milliseconds onaverage. The improvement is due to the following contributions:

- Using variable length term N-grams as items, to mitigate the effect of the skewness of the frequency distribution of unigrams. Frequently occurring sequential multi-word expressions are concatenated into one item during tokenization, resulting in transactions made up of items coming from a relatively flat distribution. Section 3.3 propose a simple method to do this; tokenizing documents (Tweets) into unigrams, then repeatedly replacing any token whose probability exceeds a threshold by two tokens formed by concatenating to it the unigrams before and after it. This simple preprocessing step results in dropping the number of itemsets mined from an hour long epoch from more than 60,000 to just above 6,000 itemsets. The reduction comes from avoiding the creation of itemsets that are combinations of function words. This preprocessing step was successfully used to improve results from the MTV algorithm [57] as well as our algorithm.

- The use of the *distinct* condition to select only itemsets which are distinctly different from their closed subsets. The condition strengthens the closure property such that an itemset is included in the mining results only if its support is different from the support of its longest subset by a substantial proportion. The closed property is easily satisfied by itemsets that are trivially different from their subsets, which results in many uninteresting itemsets – specially short itemsets made up of a commonly used term along with another term that does not have a strong association with it. Such itemsets are filtered out by the *distinct* condition, given that transactions are made up of variable length N-grams as items.

- *Distinct* itemsets are clustered into *strongly closed* itemsets for a further reduction in the number of itemsets. We propose an efficient clustering scheme that exploits the order in which itemsets are generated by LCM. The resulting clusters group together itemsets belonging to a certain fine grained topic, or an opinion within a topic, and avoids forming non-coherent clusters. The union of items in each cluster is a *strongly*

closed itemset, which groups together several *distinct* itemsets, reducing redundancy in the mining results.

Another important contribution is a formula for ranking *strongly closed* itemsets based on their temporal novelty, using the collective Information Gain of itemsets in the cluster, taking into account the dependence between itemsets in a cluster.

6.2 Results

We have shown the effectiveness of our methods by applying them to tweets posted in the time frames of various real world events. We use a minimum support threshold and an epoch length that we set according to certain characteristics of the data, which can be easily found out if the dataset is changed. However, we set the parameter that controls the distinctiveness between two itemsets, κ, to an arbitrary value according to its definition.

We have shown that the top ranked *strongly closed* itemsets from different hours are relevant to the real world events known to be taking place at those hours. For example, the top 3 itemsets from different hours in the day of the US presidential elections can be used to compose a summary of how the day unwound. We also compared our synopses for that day to the ones created by a state of the art algorithm for creating a summary based on frequent itemsets; the MTV algorithm [57]. We have shown that our algorithm efficiently creates synopses that cover the information in the summary created by MTV – itemsets that are picked by the MTV algorithm are all among the top 50 *strongly closed* itemsets. Furthermore, our synopses contain information not present in the summary created by MTV. We have also assessed the quality of the synopses against data from Google Trends. Google Trends was used to learn about real world events and entities in which Google Search users were particularly interested at different peroids of time. The synopses were then examined to see if they also reflect similar user interest. We have found that there is a strong similarity in the majority of cases, and we analyzed the cases where our synopses did not reflect the same user interest as Google Trends.

6.2.1 The Complete Results of Mining the Twitter Dataset

The complete results of mining all one-hour epochs of the Twitter dataset is available for download[1]. The mining results at different values of κ, at *minimum support* 10, are available as GZipped Tar balls. The result of each one-hour epoch in the last 3 months of 2012 (with half-hour time steps) is a tab separated file which contains: 1) the *strongly closed* itemsets, 2) their postings lists, and 3) values of several ranking formulae. The itemsets are sorted in the descending order of the ranking formula used in this thesis (column 14).

6.3 Future Work

This work provides a foundation for the use of data from social media in temporal information retrieval. The synopses we create capture people's interest at different points in times, and can be used to create a temporal profile for a query. This profile can then be used for weighting documents according to the time in which they were published.

The synopses can be also directly used for temporal query expansion. Terms from itemsets relevant to a query can be used for query expansion, thus acting as precomputed results of pseudo-revelance feedback. It is also possible to use terms from documents where relevant itemsets occur as the resultset to be used as input for pseudo-revelance feedback. The ranks of itemsets can be used to weight expansion terms, if temporal novelty is desired.

The efficiency of the proposed algorithms makes it possible to use the mining results as input to other real-time systems. For example, hashtag suggestions can be offered based on association rules created from the itemsets, such that the consequent is a hashtag (this can also be done as automatic document expansion).

Another possible future direction is to use itemsets that appear as a sequence for building coherent extractive summaries of the social media stream.

We also wish to explore ways to make use of the temporal signal during mining, such as when calculating similarity during clustering.

[1] http://plg.uwaterloo.ca/~yaboulna/thesis_results/twitter_synopses/1hr+30min_ngram5-relsupp10_oct-nov-dec/

APPENDICES

Appendix A

Unigram Tokenization

The data is tokenized into unigrams using white space and punctuation marks as separators. The characters '@', '#' and '_' are allowed within a unigram, while the rest of the punctuation marks are treated as delimiters. Only Latin characters (Unicode code points less than 'u024F') and numbers are allowed within a unigram. Any other characters that are not delimiters are skipped. If a unigram is a number then the dot and the comma characters are allowed within it. The tokenizer also does the following:

- All URLs (from 'http[s]:' to the next whitespace) are replaced by the unigram "URL".
- Runs of the same character are reduced to only 3 repetitions (for example, "coooool" is replaced by "coool").
- Hashtags are stored twice, with and without the the '#' character. The tag without the '#' character is left at tag's positions, while the other is appended at the end of the tweet. This handles cases where the hashtag is used in place of a word, such as "president #obama...".
- The apostrophe used in contractions (can't, don't, ..etc) is removed from the unigram, but does not act as a delimiter.

References

[1] Younos Aboulnaga and Charles L. A. Clarke. Frequent itemset mining for query expansion in microblog ad-hoc search. In *Proceedings of the 21st TREC Conference*, Text Retrieval Evaluation Conference (TREC), Gaithersburg, MD, USA, 2012.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.

[4] Mubarak Albathan, Yuefeng Li, and Abdulmohsen Algarni. Using patterns co-occurrence matrix for cleaning closed sequential patterns for text mining. In *2012 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 201–205. IEEE, 2012.

[5] James Allan. *Topic detection and tracking: event-based information organization*, volume 12. Kluwer Academic Pub, 2002.

[6] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study final report. 1998.

[7] James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of Topic Detection and Tracking Workshop (TDT-3)*, pages 167–174. Vienna, VA, 2000.

[8] Roberto J Bayardo, Yiming Ma, and Ramakrishnan Srikant. Scaling up all pairs similarity search. In *Proceedings of the 16th international conference on World Wide Web*, pages 131–140. Citeseer, 2007.

[9] Roberto J Bayardo Jr and Rakesh Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM, 1999.

[10] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011.

[11] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining*, pages 443–448, 2007.

[12] Jean-Francois Boulicaut, Artur Bykowski, and Christophe Rigotti. Approximation of frequency queries by means of free-sets. In *Principles of Data Mining and Knowledge Discovery*, pages 75–85. Springer, 2000.

[13] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1):5–22, 2003.

[14] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[15] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[16] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. In *ACM SIGMOD Record*, volume 26, pages 265–276. ACM, 1997.

[17] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *Principles of Data Mining and Knowledge Discovery*, pages 74–86. Springer, 2002.

[18] Toon Calders and Bart Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1):171–206, 2007.

[19] Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 66–73, 2011.

[20] Jaeho Choi and W Bruce Croft. Temporal models for microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2491–2494. ACM, 2012.

[21] Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D. Ullman, and Cheng Yang. Finding interesting associations without support pruning. *Knowledge and Data Engineering, IEEE Transactions on*, 13(1):64–78, 2001.

[22] Jacques JF Commandeur and Siem Jan Koopman. *An introduction to state space time series analysis*. OUP Oxford, 2007.

[23] John N Darroch and Douglas Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, 43(5):1470–1480, 1972.

[24] Miles Efron, Peter Organisciak, and Katrina Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 911–920. ACM, 2012.

[25] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. Mining data streams under block evolution. *ACM SIGKDD Explorations Newsletter*, 3(2):1–10, 2002.

[26] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[27] Chris Giannella, Jiawei Han, Jian Pei, Xifeng Yan, and Philip S Yu. Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, 212:191–212, 2003.

[28] Anna C Gilbert, Yannis Kotidis, S Muthukrishnan, and Martin J Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *Proceedings of the International Conference on Very Large Data Bases*, pages 79–88, 2001.

[29] Gösta Grahne and Jianfei Zhu. Reducing the main memory consumptions of fpmax* and fpclose. In *Proc. Workshop Frequent Item Set Mining Implementations (FIMI 2004, Brighton, UK), Aachen, Germany*. Citeseer, 2004.

[30] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

[31] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[32] Matthew Hoffman, David M Blei, and Francis Bach. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23:856–864, 2010.

[33] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[34] Yifan Hu. Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10(1):37–71, 2005.

[35] Jonathan Hurlock and Max L Wilson. Searching twitter: Separating the tweet from the chaff. *Proc. ICWSM 2011*, 2011.

[36] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[37] Szymon Jaroszewicz and Dan A Simovici. A general measure of rule interestingness. In *Principles of Data Mining and Knowledge Discovery*, pages 253–265. Springer, 2001.

[38] Szymon Jaroszewicz and Dan A Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 178–186. ACM, 2004.

[39] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[40] Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems (TOIS)*, 25(3):14, 2007.

[41] Roberto J. Bayardo Jr., Bart Goethals, and Mohammed Javeed Zaki, editors. *FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 2004*, volume 126 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2004.

[42] Won-Young Kim, Young-Koo Lee, and Jiawei Han. Ccmine: Efficient mining of confidence-closed correlated patterns. In *Advances in Knowledge Discovery and Data Mining*, pages 569–579. Springer, 2004.

[43] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[44] Marzena Kryszkiewicz. Concise representation of frequent patterns based on disjunction-free generators. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 305–312. IEEE, 2001.

[45] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[46] Cher Han Lau, YueFeng Li, and Dian Tjondronegoro. Microblog retrieval using topical features and query expansion. In *Proceedings of the 20th TREC Conference*, Text Retrieval Evaluation Conference (TREC), Gaithersburg, MD, USA, 2011.

[47] Young-Koo Lee, Won-Young Kim, Y Dora Cai, and Jiawei Han. Comine: Efficient mining of correlated patterns. In *Proceedings of the Third IEEE International Conference on Data Mining*, page 581, 2003.

[48] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.

[49] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[50] David D Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50. ACM, 1992.

[51] Yuefeng Li, Abdulmohsen Algarni, and Ning Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 753–762. ACM, 2010.

[52] Zhenmin Li, Zhifeng Chen, Sudarshan M Srinivasan, and Yuanyuan Zhou. C-miner: Mining block correlations in storage systems. In *Proceedings of the 3rd USENIX Conference on File and Storage Technologies*, volume 186. USENIX Association, 2004.

[53] Jimmy Lin and Gilad Mishne. A study of "churn" in tweets and real-time search queries. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[54] Bing Liu, Wynne Hsu, and Yiming Ma. Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 337–341. ACM, 1999.

[55] Guimei Liu, Haojun Zhang, and Limsoon Wong. Finding minimum representative pattern sets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 51–59. ACM, 2012.

[56] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Danah Boyd. The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. 2011.

[57] Michael Mampaey, Nikolaj Tatti, and Jilles Vreeken. Tell me what i need to know: succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 573–581. ACM, 2011.

[58] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pages 1155–1158. ACM, 2010.

[59] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Duplicate detection in click streams. In *Proceedings of the 14th international conference on World Wide Web*, pages 12–21. ACM, 2005.

[60] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[61] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.

[62] Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. *Proceedings of ICWSM*, pages 2–3, 2010.

[63] Edward R Omiecinski. Alternative interest measures for mining associations in databases. *Knowledge and Data Engineering, IEEE Transactions on*, 15(1):57–69, 2003.

[64] Ruchi Parikh and Kamalakar Karlapalem. Et: events from tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 613–620. International World Wide Web Conferences Steering Committee, 2013.

[65] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *Database Theory—ICDT'99*, pages 398–416. Springer, 1999.

[66] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Efficient mining of association rules using closed itemset lattices. *Information systems*, 24(1):25–46, 1999.

[67] Jian Pei, Jiawei Han, Runying Mao, et al. Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, volume 4, pages 21–30, 2000.

[68] Saša Petrović, Miles Osborne, and Victor Lavrenko. 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*, pages 181–189. Association for Computational Linguistics, 2010.

[69] Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. Extracting events and event descriptions from twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 105–106. ACM, 2011.

[70] Bruno Pôssas, Nivio Ziviani, Wagner Meira Jr, and Berthier Ribeiro-Neto. Set-based model: a new approach for information retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237. ACM, 2002.

[71] Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6. Association for Computational Linguistics, 2012.

[72] Alan Ritter, Oren Etzioni, Sam Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[73] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[74] Masakazu Seno and George Karypis. Finding frequent patterns using length-decreasing support constraints. *Data Mining and Knowledge Discovery*, 10(3):197–228, 2005.

[75] Abraham Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *Knowledge and Data Engineering, IEEE Transactions on*, 8(6):970–974, 1996.

[76] Avi Silberschatz and Alexander Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining*, pages 275–281, 1995.

[77] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.

[78] Nikolaj Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*, 17(1):57–77, 2008.

[79] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

[80] Fernando Tusell. Kalman filtering in r. *Journal of Statistical Software*, 39(2):1–27, 2011.

[81] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI 04)*, 2004.

[82] Jilles Vreeken, Matthijs Van Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1):169–214, 2011.

[83] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 730–735. ACM, 2006.

[84] Ke Wang, Yu He, and Jiawei Han. Mining frequent itemsets using support constraints. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 43–52, 2000.

[85] Ke Wang, Liu Tang, Jiawei Han, and Junqiang Liu. *Top down fp-growth for association rule mining.* Springer, 2002.

[86] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, volume 3, 2011.

[87] Sheng-Tang Wu, Yuefeng Li, and Yue Xu. Deploying approaches for pattern refinement in text mining. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1157–1161. IEEE, 2006.

[88] Sheng-Tang Wu, Yuefeng Li, Yue Xu, Binh Pham, and Phoebe Chen. Automatic pattern-taxonomy extraction for web mining. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pages 242–248. IEEE, 2004.

[89] Dong Xin, Hong Cheng, Xifeng Yan, and Jiawei Han. Extracting redundancy-aware top-k patterns. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–453. ACM, 2006.

[90] Dong Xin, Jiawei Han, Xifeng Yan, and Hong Cheng. Mining compressed frequent-pattern sets. In *Proceedings of the 31st international conference on Very large data bases*, pages 709–720. VLDB Endowment, 2005.

[91] Hui Xiong, P-N Tan, and Vipin Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 387–394. IEEE, 2003.

[92] Xifeng Yan, Hong Cheng, Jiawei Han, and Dong Xin. Summarizing itemset patterns: a profile-based approach. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 314–323. ACM, 2005.

[93] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 370–378. ACM, 2012.

[94] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study of retrospective and online event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.

[95] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM, 2009.

[96] Mohammed J Zaki and Ching-Jui Hsiao. Charm: An efficient algorithm for closed itemset mining. In *2nd SIAM international conference on data mining*, volume 15, pages 457–73, 2002.

[97] Mohammed J Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, Wei Li, et al. New algorithms for fast discovery of association rules. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, volume 20, pages 283–286, 1997.

[98] Yunyue Zhu and Dennis Shasha. Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 336–345. ACM, 2003.