Various Techniques for Efficient Retrieval of Contents across Social Networks Based On Events

S.Aarif Ahamed[1]
First Year M.E (CSE)
Department of CSE
M.I.E.T. EC
Tiruchirapalli
ahamed.aarif@yahoo.com

B.A.Vishnupriya[1]
First Year M.E (CSE)
Department of CSE
M.A.R. CET
Tiruchirapalli
vishnupriya907@gmail.com

V.Venkateshwaradevi[1]
Assistant Professor
Department of IT
OEC
Tiruchirapalli
devivaradaraj@gmail.com

ABSTRACT

Social Media Site (Facebook, Hi5, Last.FM, YouTube, FlickrDigg, Propeller, Reddit) is a bookmarking and community sites wherein users can share and exchange information on a wide variety of real-world events. These events range from popular, widely known ones (e.g., a concert by a popular music band) to smaller scale, local events (e.g., a local social gathering, a protest, or an accident). Some of these "event messages" might contain interesting and useful information (e.g., event time, location, participants, and opinions), others might provide little value (e.g., using heavy slang, incomprehensible language) to people interested in learning about an event. For example, a search for [*PSLV-C21* rocket Launch] on YouTube returns over 30,000 videos. Even smaller events often feature dozens to hundreds of different content items. It is most important to select and prioritize the event content to avoid overwhelming the users with too much information. This paper, explore approaches for finding representative messages among a set of messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information. A comprehensive survey and study of various approaches for retrieving high quality, relevant event content has been presented in this paper.

Keywords

Social Media Site, Event Content, Content Retrieval, User Interactions, Clustering, Classification

1. INTRODUCTION

The ease of publishing content on social media sites brings to the Web an ever increasing amount of content captured during and associated with real-world events. Sites like Flickr, YouTube, Facebook and others host user-contributed content for a wide variety of events. These range from widely known events, such as presidential inaugurations, to smaller, community-specific events, such as annual conventions and local gatherings. Event-based information sharing and seeking are common user interaction scenarios on the Web today. The bulk of information from events is contributed by individuals through social media channels: on photo and video-sharing sites (e.g., Flickr, YouTube), as well as on social networking sites (e.g., Facebook, Twitter). This event-related information can appear in many forms, including status updates in anticipation of an event, photos and videos captured before, during, and after the event, and messages containing post event reactions. Importantly, for known and upcoming events (e.g., concerts, parades, and conferences) revealing, structured information (e.g., title, description, time, and location) is often explicitly available on user-contributed event aggregation platforms (e.g., Last.fm events, EventBrite, Facebook events).

The paper is organized as follows. In section 2, various techniques for identifying high quality, relevant messages from social networks will be introduced. Finally, section 3 is a conclusion.

2. LITERATURE SURVEY

2.1 Finding High-Quality Content in Social Media

Social media in provides a rich variety of information sources (content and an array of non-content information's, such as links between items and explicit quality ratings from members of the community). The task of identifying high-quality content in sites based on user contributions on social media sites becomes increasingly important. Eugene Agichtein et al [3] introduce a general classification framework for exploiting such community feedback to automatically identify high quality content.

2.1.1 Steps involved in finding high-quality content in social media:

Step 1: Intrinsic content quality

Intrinsic quality metrics refers to quality of the content of each item that can be categorized by means of semantic features organized as follows:

1. **Punctuation and typos:** Poor quality text, and particularly of the type found in online sources, is often marked with low conformance to common writing practices. For example, capitalization rules may be ignored; excessive punctuation particularly repeated ellipsis and question marks may be used, or spacing may be irregular.
2. **Syntactic and semantic complexity:** Advancing from the punctuation level to more involved layers of the text, other features in this subset quantify the syntactic and semantic complexity of it. These include simple proxies for complexity such as the average number of syllables per word or the entropy of word lengths, as well as more intricate ones such as the readability measures.
3. **Grammaticality:** Finally, to measure the grammatical quality of the text. Some part-of-speech sequences are typical of correctly formed questions: e.g., the sequence\whenjhowjwhy to (verb)" (as in \how to identify. . . ") is typical of lower-quality questions, whereas the sequence\whenjhowjwhy (verb) (personal pronoun) (verb)" (as in \how do I remove. . . ") is more typical of correctly-formed content.

Step 2: User relationships

The interactions of users are organized around questions: the main forms of interaction among the users are (i) asking a question, (ii) answering a question, (iii) selecting best answer, and (iv) voting on an answer.

Step 3: Usage statistics

Readers of the content (who may or may not also be contributors) provide valuable information about the items they find interesting. In particular, usage statistics such as the number of clicks on the item and dwell time have been shown useful in the context of identifying high quality web search results, and are complementary to link-analysis based methods.

For example, all items within a popular category such as celebrity images or popular culture topics may receive orders of magnitude more clicks than, for instance, science topics. Nevertheless, when normalized by the item category, the deviation from expected number of clicks can be used to infer quality directly, or can be incorporated into the classification framework.

Step 4: Classification Framework

A sequence of decision trees [1] [2] is constructed so that each tree minimizes the error on the residuals of the preceding sequence of trees; a stochastic element is added by randomly sampling the data repeatedly before each tree construction, to prevent over fitting. A particularly useful aspect of boosted trees for our settings is their ability to utilize combinations of sparse and dense features. Given a set of human-labeled quality judgments, the classifier is trained on all available features, combining evidence from semantic, user relationship, and content usage sources. The judgments are tuned for the particular goal. For example, we could use this framework to classify questions by genre or asker expertise.

2.2 Selecting Quality Twitter Content for Events

Hila Becker et al [7] finding representative messages among a set of Twitter messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information.

This can be achieved with two concrete steps. First, identify each event—and its associated Twitter messages—using an online clustering technique that groups together topically similar Twitter messages. Second, for each identified event cluster, select messages that best represent the event.

2.2.1 Steps involved in retrieval of quality twitter contents for an event:

Step 1: Identifying Event Content

Associate Twitter messages with events using an online clustering framework. Specifically, use an incremental, online clustering algorithm to effectively cluster a stream of Twitter messages in a scalable fashion, without requiring a priori knowledge of the number of clusters. These features of the clustering algorithm are particularly desirable for this domain since Twitter messages are constantly produced and new events are added to the stream over time. Figure 1 describes the clustering process.

The weights are assigned to each cluster during a supervised training phase, and used to determine each cluster's influence on the overall ensemble similarity assignment [5]. By assigning a weight to a cluster, which provides how successful the cluster was in capturing document similarity on a training set, and therefore how likely it is to correctly indicate the similarity of unseen document pairs.

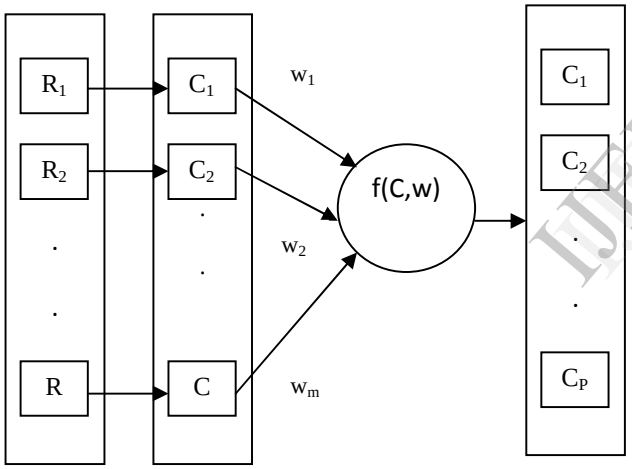

Fig 1: Clustering Process

Step 2: Event Content Selection

Once the events and their associated Twitter messages are identified, then comes the selection of a subset of these messages for presentation. Selection of messages for each identified event with three desired attributes: quality, relevance, and usefulness.

- **Quality** refers to the textual quality of the messages, which reflects how well they can be understood by a human. High-quality messages contain crisp, clear, and effective text that is easy to understand. Low-quality messages, on the other hand, contain incomprehensible text, heavy use of short-hand notation, spelling and grammatical errors, and types. Interestingly, the quality of a message is largely independent of its associated event.
- **Relevance** refers to how well a Twitter message reflects information related to its associated event. Highly relevant messages clearly refer to or describe their associated event. Messages are not relevant to an event if they do not refer to the event in any way.
- **Usefulness** refers to the potential value of a Twitter message for someone who is interested in learning details about an event. Useful messages should provide some insight about the event, beyond simply stating that the event occurred. The level of usefulness of Twitter messages varies. Messages that are clearly useful provide potentially interesting details about the event. Messages that are clearly not useful provide no context or information about the event. Other messages may reflect a user's opinion about the event, where somewhat useful event information is directly stated or can be inferred.

2.3 Identifying Content for Planned Events across Social Media Sites:

Hila Becker et al [4] extended the previously discussed approach there by explicitly providing event features such as title (e.g., \ *PSLV-C21* rocket Launch), description, time/date, location, and venue to automatically formulate queries used to retrieve related social media content from multiple social media sites and identify the top-k such documents from each site, according to given site-specific scoring functions. In order to achieve this, generate a variety of queries for each event to collectively retrieve matching social media documents from multiple sites. Since each event could potentially have many associated social media documents, filtering process is carried out to filter the set of documents to the top-k most similar documents, using given site-specific scoring functions. As a result the top-k most similar documents are presented to a user. Figure 2 presents an overview of query generation approach.

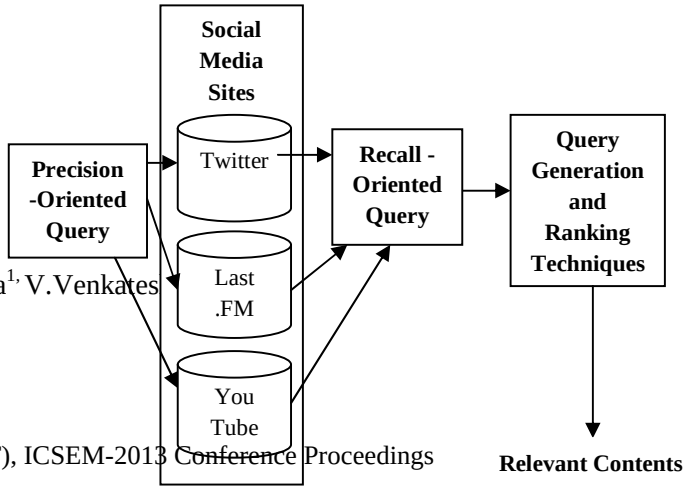

Fig 2: Query Generation

2.3.1 Steps involved in retrieval of related social media content from multiple social media sites:

Step 1: Precision-Oriented Query Building Strategies

First step towards retrieving social media documents for planned events consists of simple query generation strategies that are aimed to collectively retrieve a set of social media documents with high-precision results. The precision-oriented queries for an event consist of combinations of one or more event features (e.g., title, time/date and venue) [4].

Step 2: Recall - Oriented Query Building Strategies

The precision-oriented queries return high-precision social media documents for an event; the number of these high-precision documents is generally low.

To improve recall achieved in the first step, term extraction and frequency analysis techniques are used on the high-precision results to generate recall-oriented queries and retrieve additional documents for the event. Event's title, description, and any retrieved results from the precision-oriented techniques are treated as "ground-truth" data for the event.

Using the ground-truth data for each event, query formulation techniques was designed to capture terms that uniquely identify each event. These terms should ideally appear in any social media document associated with the event but also be broad enough to match a larger set of documents than possible with the precision-oriented queries.

Recall-oriented queries can be selected in two steps. First, generation of a large set of candidate queries for each event using two different term analysis and extraction techniques. Then, to select the most promising queries out of a potentially large set of candidates, a variety of query ranking strategies were explored and identify the top queries according to each strategy.

(1) Term Analysis:

The first query candidate generation technique aims to extract the most frequently used terms, while weighing down terms that are naturally common in the English language. To select these terms, we compute term frequencies over the ground-truth data for word unigrams, bigrams, and trigrams. Then eliminate stop words and remove infrequent n- grams (determined automatically based on the size of the ground-truth corpus).

(2) **Term Extraction**:

The second query candidate generation technique aims to identify meaningful event-related concepts in the ground-truth data using an external reference corpus. For this, a Web-based term extractor over our available textual event data [4] is used. This term extractor leverages a large collection of Web documents and query logs to construct an entity dictionary, and uses it along with statistical and linguistic analysis methodologies to find a list of significant terms. The extracted terms for each event serve as additional recall-oriented query candidates, along with the term-frequency query candidates described above.

Step 3: Query Generation and Ranking Techniques

Each of the techniques described above could potentially generate a large set of candidate queries. However, many of these queries could be noisy, too general, or describing a specific or non - central aspect of the event. Issuing hundreds of queries for each event is not scalable and could potentially introduce substantial noise, so it is important to further reduce the set of queries to the most promising candidates. A variety of strategies were explored for selecting the top candidate queries out of all possible queries for each event. There two basic options to rank the queries for selection, namely, using

(1) the "specificity "of the queries, as determined by the n-gram score on the Microsoft Web document corpus
(2) Variations of a "temporal" profile of the queries, determined by analyzing the volume of matching documents for the queries over time.

Each alternative technique selects the top-10 queries according to the associated ranking criterion, as follows:

- **MS n-gram Score (MS):** n-gram score of the query from the Microsoft Web n-gram Service

- **Time Ratio (TR):** ratio of the number of documents created in the 48 hours before and after the event to the number of documents created in the week before and after the event
- **Restricted Time Ratio (RTR):** ratio of the number of documents created in the 24 hours before and after the event to the number of documents created in the week before and after the event
- **MS n-gram Score and Time Ratio (MS-TR):** MS score multiplied by TR score
- **MS n-gram Score and Restricted Time Ratio (MS-RTR):** MS score multiplied by RTR score

3. CONCLUSION

Social media sites have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events. These events range from popular, widely known ones to smaller scale, local events. Some of these event messages" might contain interesting and useful information others might provide little value .It is most important to select and prioritize the event content to avoid overwhelming the users with too much information. This paper, explore approaches for finding representative messages among a set of messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information. The extensive study of these three techniques reveals that "Identifying Content for Planned Events across Social Media Sites" proves more profitable & efficient retrieval of high quality relevant event contents from various social media sites than the other two techniques. The future work includes retrieval of relevant images in addition to the contents from various social media sites.

4. ACKNOWLEDGMENTS

We thanks to the light, our god, who guided us through the way. We have taken efforts in this paper. However, it would not have been possible without his kind support and help of many individuals and organizations. I wish to avail myself of this opportunity, express a sense of gratitude and love to my friends and my beloved parents for their manual support, strength, help and for everything.

5. REFERENCES

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, Special Issue on Learning and Discovery in Knowledge-Based Databases.

[2] Han and Kamber "Data Mining: Concepts and Techniques", Mogan Kaufmann Publishers.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM '08), 2008.

[4] Hila Becker_y, Dan Itery, Mor Naamanz, Luis Gravano: Identifying Content for Planned Events Across Social Media Sites.

[5] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10), 2010.

[6] H. Becker, M. Naaman, and L. 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.

[7] H. Becker, M. Naaman, and L. Gravano. Selecting quality Twitter content for events. In proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11), 2011.

[8] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11), 2011.