Lecture Notes in Social Networks

Mehmet Kaya
Özcan Erdoğan
Jon Rokne *Editors*

From Social Data Mining and Analysis to Prediction and Community Detection

Springer

Lecture Notes in Social Networks

Series editors

Reda Alhajj, University of Calgary, Calgary, AB, Canada
Uwe Glässer, Simon Fraser University, Burnaby, BC, Canada

Advisory Board

Charu C. Aggarwal, IBM T.J. Watson Research Center, Hawthorne, NY, USA
Patricia L. Brantingham, Simon Fraser University, Burnaby, BC, Canada
Thilo Gross, University of Bristol, Bristol, UK
Jiawei Han, University of Illinois at Urbana-Champaign, IL, USA
Huan Liu, Arizona State University, Tempe, AZ, USA
Raul Manasevich, University of Chile, Santiago, Chile
Anthony J. Masys, Centre for Security Science, Ottawa, ON, Canada
Carlo Morselli, University of Montreal, QC, Canada
Rafael Wittek, University of Groningen, The Netherlands
Daniel Zeng, The University of Arizona, Tucson, AZ, USA

More information about this series at http://www.springer.com/series/8768

Mehmet Kaya • Özcan Erdoğan • Jon Rokne
Editors

From Social Data Mining and Analysis to Prediction and Community Detection

Springer

Editors

Mehmet Kaya
Department of Computer Engineering
Firat University
Elazig, Turkey

Özcan Erdoğan
Ministry of Interior
Ankara, Turkey

Jon Rokne
Department of Computer Science
University of Calgary
Calgary, AB, Canada

ISSN 2190-5428 ISSN 2190-5436 (electronic)
Lecture Notes in Social Networks
ISBN 978-3-319-51366-9 ISBN 978-3-319-51367-6 (eBook)
DOI 10.1007/978-3-319-51367-6

Library of Congress Control Number: 2017931681

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Introduction

This volume is a compilation of the best papers presented at the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'2015), held in Paris, France, August 2015. The authors of these papers were asked to provide extended versions of the papers that were then subjected to an additional refereeing process. Within the broader context of online social networks, the volume focuses on important and upcoming topics such as attempting to understand the context of messages propagating in social networks, classifying sentiments and defining and understanding local communities.

From Social Data Mining and Analysis to Prediction and Community Detection

The importance of social networks in today's society cannot be underestimated. Millions of individuals use social networks to communicate with friends every day, and for many people (especially young ones) the social network presence is of vital importance for their identity.

Social networks are a rich source of information, and several of the papers in this volume focus on how to extract this information. The information gleaned from social networks can inform businesses, governments, social agencies, cultural groups and so on.

Social networks lead to questions such as privacy, ethics and data ownership. These questions and other issues are a fertile ground for further research.

The first paper in this volume is "An Offline-Online Visual Framework for Clustering Memes in Social Media" by Anh Dang, Abidalrahman Moh'd, Anatoliy Gruzd, Evangelos Milios and Rosane Minghim. The paper discusses the dissemination and clustering of memes in social networks where memes are "an element of a culture or system of behaviour that may be considered to be passed from one individual to another by non-genetic means, especially imitation or a humorous image, video, piece of text, etc., that is copied and spread rapidly by Internet users" (Bing definition). It turns out that social networks are ideal for the dissemination of memes and it is therefore interesting to study how the memes propagate through such networks and how the memes cluster in the networks. The authors compute similarity scores between texts containing memes and assess cluster memberships depending on the scores. The social network Reddit is studied in detail, and the clustering of memes in this network is used to detect emerging events. Experimental results show that their method implemented using the Google Trigram Method provides reasonable results.

Given that more than 200 billion e-mail messages are sent each day, it is to be expected that some messages will be sent to the wrong address. Most of us are likely guilty of doing it. Sometimes this can have serious consequences especially if sensitive information is transmitted. The paper by Zvi Sofershtein and Sara Cohen entitled "A System for Email Recipient Prediction" effectively predicts e-mail recipients when an e-mail history is available. This system takes into account a variety of clues from e-mail histories to predict recipients when there is sufficient data. The paper proposes a system based on a set of features and assesses the contribution to the precision of the system for each of the features. The system is tested on data sets and on various domains such as the Enron data set, a political data, a Gmail-English data set, a Gmail-Hebrew data set and a combined data set. Properly applied their system can reduce the number of unfortunate misdirected messages and increase the speed by which correctly addressed e-mails are composed.

Authenticating and verifying content of messages is common to many activities in today's world. Verified messages can be beneficial, whereas messages containing false information can be harmful (witness the scandal surrounding the falsified messages relating to the health benefits of sugar versus fat). Both true and false messages can be disseminated using online social networks. In paper "A Credibility Assessment Model for Online Social Network Content", Majed Alrubaian, Muhammad Al-Qurishi, Mabrook Al-Rakhami and Atif Alamri present an algorithmic approach for assessing the credibility of Twitter messages. They report encouraging results being able to achieve accuracies in the 80% range on two specific data sets.

Whereas the previous paper assessed the credibility of Twitter messages in Arabic, the paper by Mohammed Bekkali and Abdelmonaime Lachkar entitled "Web Search Engine Based Representation for Arabic Tweets Categorization" attempts to enhance Tweets (or in general short length messages) composed in Arabic by providing contextual information. They note that there is scant research in machine understanding of Arabic, partly due to its structure being different from the more studied European languages, and the understanding of short messages in Twitter in Arabic is therefore a twofold problem of both encoding (i.e. language) and the choice of language itself.

Preserving and increasing capital by investing in stocks is fraught with uncertainty. Reducing this uncertainty by applying sentiment analysis is discussed in the paper “Sentiment Trends and Classifying Stocks Using P-Trees” by Arijit Chatterjee and William Perrizo. The authors mine the large volume of Tweets generated each day for sentiment trends for specific stock symbols. They use Twitter API and P-trees to gain insight into what ticker symbols are “hot” and hence what ticker symbols and businesses should be a reasonable investment. Stock prices and sentiments are graphed for the two stock symbols AAPL (Apple) and FB (Facebook), and they conclude that the trends of stock prices follow the movements of the sentiments in a general sense.

Social networks in their simplest form consist of nodes and links between nodes. Essentially the links are the minimal information needed to form such networks. If the nodes contain richer information, then further insights can be gained from the networks. In their paper “Mining Community Structure with Node Embeddings”, Thuy Vu and D. Stott Parker show that embeddings can reflect community structure. Node embeddings were implemented for the DLBP network, and it was shown that they were useful in finding communities.

Finding local communities in a social network has a number of applications. In the paper “A LexDFS-Based Approach on Finding Compact Communities”, Jean Creusefond, Thomas Largillier and Sylvain Peyronnet apply an efficient graph transversal algorithm to identify clusters and hence communities in a social network. The paper proposes a new measure for connectedness. In their definition a subset of the social network is said to be connected if the distances between the nodes in the subset are small. They apply the LexDFS algorithm to compute connected communities and compare the results with six previously developed clustering algorithms.

Societies need laws and regulations to function effectively and to limit the damage unhinged individuals and entities can do if left unchecked. Laws and regulations therefore have a long history dating back to, at least, the Babylonians. Past laws and regulations are often modified and built upon to take into account changes and new advances in society, such as the Internet. Especially due to the Internet, the laws pertaining to privacy and data ownership have still to be defined in an acceptable form. In the past creating and modifying laws and regulations was the domain of well-paid legal experts. As society has become more and more complex and with the global reach of modern economies, the understanding and applying of laws has become more and more difficult and informatics tools are now applied to legal work. The financial sector is a fundamental and complex part of modern society. While it is the focus of much criticism, it is also an essential component of modern society. The aim of laws and regulations for the financial sector is to ensure a smooth functioning of economies while limiting excesses such as gambling with toxic derivatives (and collection obscene rewards for doing so). The resulting corpus of laws and regulations is very large and complex (especially in the United States of America), and it is difficult to get an overview of the corpus even for the experts. The paper “Computational Data Sciences and the Regulation of Banking and Financial Services” by Sharyn O’Halloran, Marion Dumas, Sameer Maskey, Geraldine McAllister and David Park attempts to alleviate some of the heavy lifting in this area by applying sophisticated computational algorithms to the corpus.

The final paper of the volume "Frequent and Non-frequent Sequential Itemsets Detection" by Konstantinos F. Xylogiannopoulos, Panagiotis Karampelas and Reda Alhajj considers the problem of detecting repeated patterns in a string. These patterns are classified as frequent sequential itemsets if they pertain to sets of events that are ordered. The proposed algorithm is based on transforming a set of transactions into a suitable data structure that then is processed using a novel process. This process was used for experiments with two data sets. The first data set was the "Retail" data set consisting of 88,162 transactions, and the second data set was the "Kosavak" data set with 990,002 transactions. In both cases all the frequent data sets were detected regardless of support value.

To conclude this preface, we would like to thank the authors who submitted papers and the reviewers who provided detailed constructive reports which improved the quality of the papers. Various people from Springer deserve large credit for their help and support in all the issues related to publishing this book.

Elazig, Turkey
Ankara, Turkey
Calgary, AB, Canada
October 2016

Mehmet Kaya
Özcan Erdoğan
Jon Rokne

Contents

An Offline–Online Visual Framework for Clustering Memes in Social Media

Anh Dang, Abidalrahman Moh'd, Anatoliy Gruzd, Evangelos Milios, and Rosane Minghim

1 Introduction

Online Social Networks (OSNs) are networks of online interactions and relationships that are formed and maintained through various social networking sites such as Facebook, LinkedIn, Reddit, and Twitter. Nowadays, hundreds of millions of people and organizations turn to OSNs to interact with one another, share information, and connect with friends and strangers. OSNs have been especially useful for disseminating information in the context of political campaigning, news reporting, marketing, and entertainment [32].

OSNs have been recently used as an effective source for end users to know about breaking-news or emerging memes. A meme is a unit of information that can be passed from person to person in OSNs [29]. Despite their usefulness and popularity, OSNs also have a "negative" side. As well as spreading credible information, OSNs can also spread rumours, which are truth-unverifiable statements. For example, so many rumour-driven memes about swine flu outbreak (e.g., "swine flu pandemic meme" in Fig. 1) were communicated via OSNs in 2009 that the US government

A. Dang (✉) • A. Moh'd • E. Milios
Faculty of Computer Science, Dalhousie University, 6050 University Avenue, PO BOX 15000, Halifax, NS, Canada B3H 4R2
e-mail: anh@cs.dal.ca; amohd@cs.dal.ca; eem@cs.dal.ca

A. Gruzd
Ted Rogers School of Management, Ryerson University, 55 Dundas Street West, Toronto, ON, Canada M5G 2C3
e-mail: gruzd@ryerson.ca

R. Minghim
University of São Paulo-USP, ICMC, São Carlos, Brazil
e-mail: rminghim@icmc.usp.br

© Springer International Publishing AG 2017

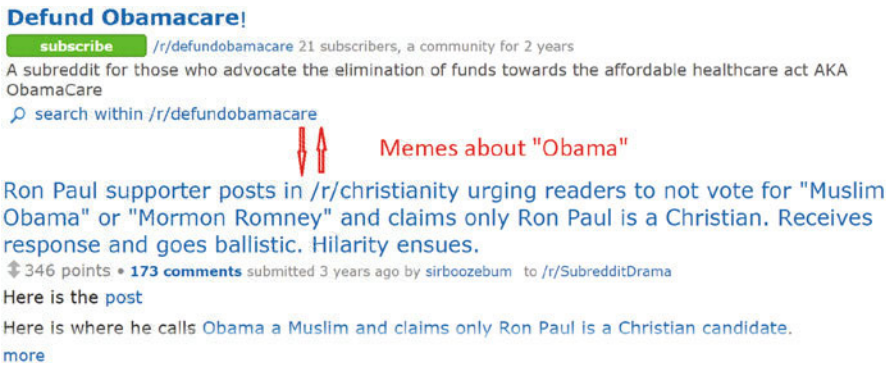
M. Kaya et al. (eds.), *From Social Data Mining and Analysis to Prediction and Community Detection*, Lecture Notes in Social Networks,
DOI 10.1007/978-3-319-51367-6_1

Fig. 1 A word cloud example of popular memes in OSNs

had to tackle it officially on their website [22, 37]. Problems like these (i.e., rumour-driven memes going viral) are unfortunately not isolated and prompt the question of how to identify and limit the spread of rumours in OSNs. In order to detect rumours, we have to identify memes that are rumour-related in OSNs. Clustering is a simple and efficient unsupervised process to identify memes in OSNs by grouping similar information into the same category. However, traditional clustering algorithms do not work effectively in OSNs due to the heterogeneous nature of social network data [31]. Labelling massive amounts of social network data is an intensive task for classification. To overcome these limitations, this paper proposes a semi-supervised approach with relevance user feedback for detecting the spread of memes in OSNs.

In text clustering, a similarity measure is a function that assigns a score to a pair of texts in a corpus that shows how similar the two texts are. Computing similarity scores between texts is one of the most computationally intensive and important steps for producing a good clustering result [6]. For a meme clustering task, this process is usually hindered by the lack of significant amounts of textual content, which is an intrinsic characteristic of OSNs [17, 27, 31]. For example, in Reddit.com, most submission titles are very short and concise. Although the title of a submission may provide meaningful information about the topic, the titles may not provide enough information to determine if two submissions are discussing the same topic. In Fig. 2, two Reddit submissions are both talking about "Obama," but one is discussing the meme "Obamacare," while the other is discussing the rumour-related meme "Obama is a Muslim." The sparsity of Reddit submission title texts significantly contributes to the poor performance of traditional text clustering techniques for grouping submissions into the same category. We, therefore, propose strategies to leverage the use of references to external content.

A submission may include one or more comments from users, which discuss the submission topic. It can also contain a URL that points to an external article that further discusses the topic of the submission. Similarly, a submission may include an image that also provides more valuable information about the submission topic. By introducing the use of comments, URL content, and image content of a submission, we exploit more valuable data for text clustering tasks, which helps detect memes in OSNs more efficiently.



Defund Obamacare!

/r/defundobamacare 21 subscribers, a community for 2 years

A subreddit for those who advocate the elimination of funds towards the affordable healthcare act AKA ObamaCare

Ron Paul supporter posts in /r/christianity urging readers to not vote for "Muslim Obama" or "Mormon Romney" and claims only Ron Paul is a Christian. Receives response and goes ballistic. Hilarity ensues.

346 points • 173 comments submitted 3 years ago by sirboozebum to /r/SubredditDrama

Here is the post

Here is where he calls Obama a Muslim and claims only Ron Paul is a Christian candidate.

Fig. 2 Reddit submissions about the same meme "Obama." The *top submission* discusses the meme "Obamacare." while the *bottom submission* discusses the meme "Obama is a Muslim"

Vector space models are commonly used to represent texts for a clustering task. In these models, each text is represented as a vector where each element corresponds to an attribute extracted from the text. One of the benefits of these models is their simplicity in calculating the similarity between two vectors based on linear algebra. The two most famous models are TF-IDF (Term Frequency—Inverse Term Frequency) and Bag-of-Words. However, those models rely solely on lexical representation of the texts, which does not capture semantic relatedness between words in a text. For example, the use of polysemy and synonymy are very popular in several types of texts and play an important role in determining whether two words, concepts or texts are semantically similar. This motivates many researchers to explore the advantage of semantic similarity in the task of text clustering by utilizing word relatedness through external thesauruses like WordNet and Wikipedia [28, 38]. However, they remain far from covering every word and concept used in OSNs. This paper explores Google n-grams algorithm of Islam et al. [30] which uses Google n-grams dataset to compute the relatedness between words for computing similarity scores, and proposes two novel strategies to combine those scores for the task of clustering memes.

Although clustering streaming and time series data are established fields in the area of text clustering [14, 23–25, 41], clustering memes in OSNs has just started to gain attention recently [2, 3, 31, 42]. OSN data has both characteristics of streaming and time series data, as well as another important characteristic. The volume of OSN data is massive and cannot be handled efficiently by traditional streaming and time series clustering algorithms [31]. In order to tackle that problem, we propose a novel approach to speed up the processing of online meme clustering that uses both semantic similarity and Wikipedia concepts to efficiently store and summarize OSN data in real time.

With the increasing amount of online social network data, understanding and analyzing them is becoming more challenging. Researchers have started to employ human's ability to effectively gain visual insight on data analysis tasks. The task of clustering memes shares some similarity with clustering text, but they are also intrinsically different. For example, social network data is usually poorly written and content-limited. This reduces the quality of clustering results. For a Reddit submission, the relationships between the title, comment, image, and URL sometimes are disconnected (e.g., a title has a different subject from the content). In this paper, we developed a visualization prototype to allow users to better distinguish the similarity between submissions and use this feedback to improve the clustering results.

This paper extends previous work of Dang et al. [18] by formalizing the problem of meme clustering and proposes a novel approach for clustering Reddit submissions. It makes the following contributions:

- Extends and improves the similarity scores between different elements of Reddit submission of Dang et al. [18] by introducing the use of Wikipedia concepts as an external knowledge.
- Introduces a modified version of Jaccard Coefficient that employs the use of text semantic similarity when comparing the similarity score between two sets of Wikipedia concepts.
- Proposes an Offline–Online clustering algorithm that exploits semantic similarity and Wikipedia concepts to achieve good clustering results in real time. The offline clustering component computes and summarizes cluster statistics to speed up the process of the online clustering component. In addition, for each cluster, we adopt the damped window model and propose a novel approach to summarize each cluster as a set of Wikipedia concepts where each concept is assigned a weight based on its recency and popularity. The online clustering component applies a semantic version of Jaccard Coefficient.
- The experiments show the use of Wikipedia concepts increases the accuracy result of the meme clustering tasks. Although only using Wikipedia concepts as a similarity score does not increase the clustering result, using both Wikipedia concepts and text semantic similarity increase the clustering accuracy for both offline and online clustering components.

2 Related Work

This section presents current research on text semantic similarity and detecting the spread of memes in OSNs.

2.1 Similarity Measures and Text Clustering

Several similarity measures have been proposed in the literature for the task of text clustering. The most popular ones are lexical measures like Euclidean, Cosine, Pearson Correlation, and Extended Jaccard measures. Strehl et al. [40] provided a comprehensive study on using different clustering algorithms with these people measures. The authors used several clustering algorithms on the YAHOO dataset, and showed that Extended Jaccard and Cosine similarity performed better and achieved results that are close to a human-labelling process. However, lexical similarity measures do not consider the semantic similarity between words in the texts.

Some researchers have taken advantage of the semantic relatedness of texts by using external resources to enrich word representation. In [38], the authors suggested using WordNet as a knowledge base to determine the semantic similarity between words. The experiment results have shown that external knowledge bases like WordNet improve the clustering results in comparison to the Bag-of-Words models. Hu et al. [28] proposed the use of Wikipedia as external knowledge for text clustering. The authors tried to match concepts in text into Wikipedia concepts and categories. Similarity scores between concepts are calculated based on the text content information, as well as Wikipedia concepts, and categories. The experiment results have shown that using Wikipedia as external knowledge provided a better result than using WordNet due to the limited coverage of WordNet. Bollegala et al. [8] proposed the use of information available on the Web to compute text semantic similarity by exploiting page counts and text snippets returned by a search engine. Our work is intuitively different from these approaches, as it introduces the use of word relatedness based on the Google n-grams dataset [9]. The proposed semantic similarity scores between texts are calculated based on that algorithm to handle the low quality (i.e., poor writing) of social network data. Using Google n-grams dataset as external knowledge is more effective than textual as well as other semantic approaches, as the Google n-grams dataset has more coverage than other semantic approaches.

2.2 Online Clustering Algorithms in OSNs

This section discusses the related work of event detection and online meme clustering in OSNs. The proposed online meme clustering algorithm takes advantage of the current work of both clustering streaming data and clustering time series data. Clustering streaming data has been actively researched in the literature. Aggarwal et al. [3] proposed a graph-based sketch structure to maintain a large number of edges and nodes at the cost of potential loss of accuracy. Zhao and Yu [42] extended the graph-based clustering streaming algorithms with side information, such as user metadata in OSNs. As OSN data is dependent on its temporal context, time series is another important feature of clustering streaming data algorithms. We present related work of the three types of time series clustering algorithms in the literature: (1) landmark window approach, (2) sliding window approach, and (3) damped window approach [39]. In clustering streaming data, landmark-based models consider all the historical data from a landmark time and all data have an equal weight [11, 24, 36]. Sliding-based models are common stream processing models which only examine data at a fixed-time window (e.g., last 5 min or last 24 h) [14, 25, 35]. Damped window models introduced the use of decay variable to replace old data to increase the accuracy of streaming clustering results [13, 23]. JafariAsbagh et al. [31] used a sliding window approach for detecting memes in real time that does not consider the topic evolution and persistence. As the spread of memes in OSNs is dependent on the meme topics and its context [26], the proposed online meme clustering algorithm explores the damped window approach which considers the frequency and recency of memes. Researchers also investigate if the use of external knowledge (e.g., Wikipedia) helps the clustering results for social media texts. Banerjee et al. [4] introduced the use of Wikipedia as external knowledge to improve the accuracy results for short texts. Dang et al. [18] used text semantic similarity computed from Google n-grams dataset to alleviate the problem of shortness and noise of OSN data.

Scientists also explored the use of visualization for text clustering with relevance user feedback. Lee et al. [33] introduced iVisClustering, an interactive visualization framework based on LDA topic modelling. This system provides some interactive features, such as removing documents or clusters, moving a document from one cluster to another, merging two clusters, and influencing term weights. Choo et al. [16] presented an interactive visualization for dimension reduction and clustering for large-scale high-dimensional data. The system allows users to interactively try different dimension reduction techniques and clustering algorithms to optimize the clustering results. One of the limitations of these systems is that they focus on the clustering algorithms and results and have limited supports for combining similarity scores for different parts of a text (e.g., the title and body of a text). This paper introduces a visualization prototype to combine different similarity scores for our clustering process interactively and incrementally.

2.3 Detecting Memes in Online Social Networks

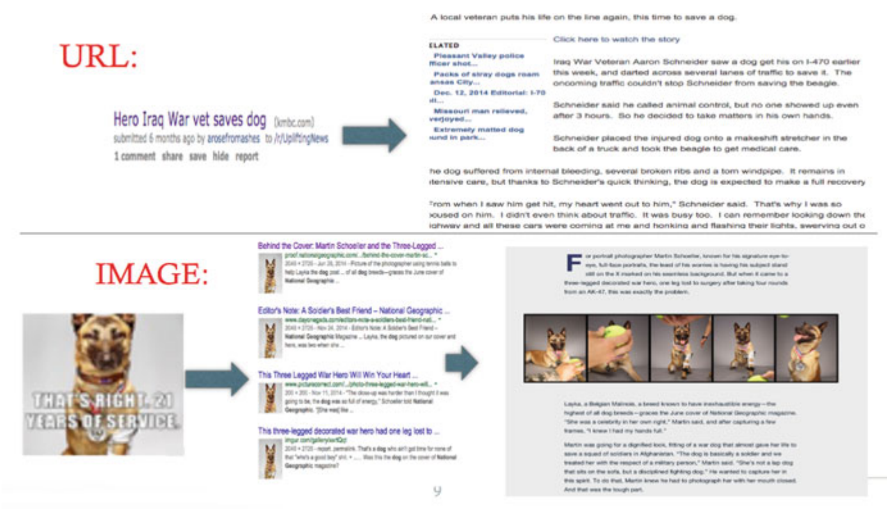
Recently, researchers have started adapting state-of-the-art clustering algorithms to OSN data. Leskovec et al. [34] proposed a meme-tracking framework to monitor memes that travel through the Web in real time. The framework studied the signature path and topic of each meme by grouping similar short, distinctive phrases together. One drawback of this framework is that it only applies lexical content similarity to detect memes. This did not work well for memes that are related but not using the same words, and those that are short and concise (e.g., Tweets on Twitter). Cataldi et al. [10] proposed an approach that monitored the real-time spread of emerging memes in Twitter. The authors defined an emerging term as one whose frequency of appearance had risen within a short period and had not emerged or was only rarely discussed in the past. A navigable topic graph is constructed to connect semantically related emerging terms. Emerging memes are extracted from this graph based on semantic relationships between terms over a specified time interval. Becker et al. [5] formulated the problem of clustering for event detection and proposed a supervised approach to classify tweets using a predefined set of features. The proposed approach includes various types of features: textual, temporal, and spatial. Aggarwal and Subbian [1] presented a clustering algorithm that exploits both content and network-based features to detect events in social streams. The proposed algorithm uses knowledge about metadata of Twitter users. Thom et al. [41] developed a system for interactive analysis of location-based microblog messages, which can assist in the detection of real-world events in real time. This approach uses X-means, a modified version of K-means, to detect emerging events. Finally, JafariAsbagh et al. [31] introduced an online meme clustering framework using the concept of Protomemes. Each Protomeme is defined based on one of the atomic information entities in Twitter: hashtags, mentions, URLs, and tweet content. An example of Protomeme is the set of tweets containing the hashtag #All4Given. This approach uses a sliding window model that can lead to good offline prediction accuracy but not suitable for online streaming environments. As online meme clustering algorithms require low prediction and training costs, our proposed online meme clustering algorithm stores cluster summary statistics using Wikipedia concepts and applies a damped window approach with Offline–Online components for clustering memes in OSNs. Although Twitter has been the most popular OSN for detecting memes, little work has been done to detect rumour-related memes on Reddit.

3 Reddit Social Network

Reddit, which claims to be "the front page of the internet," is a social news website, where users, called redditors, can create a submission or post direct links to other online content. Other redditors can comment or vote to decide the rank of this submission on the site. Reddit has many subcategories, called sub-reddits that are organized by areas of interests. The site has a large base of users who discuss a wide range of topics daily, such as politics and world events. Alexa ranks Reddit.com as the 24th most visited site globally. Each Reddit submission has the following elements:

- **Title:** The title summarizes the topic of that submission. The title text is usually very short and concise. The title may also have a description to further explain it.
- **Comments:** Users can post a comment that expresses their opinions about the corresponding submission or other user comments. Users can also vote comments up or down.
- **URL:** Each submission may contain a link to an external source of information (e.g., news articles) that is related to the submission.
- **Image:** Submissions may also have a link to an image that illustrates the topic of the submission.

Figure 3 explains how to collect image and URL content from Reddit submissions. Unlike other OSNs, Reddit is fundamentally different in that it implements an open data policy; users can query any posted data on the website. For example, other



Hero Iraq War vet saves dog (kmbc.com)
submitted 6 months ago by arosefromashes to /r/UpliftingNews
1 comment share save hide report

A local veteran puts his life on the line again, this time to save a dog.

ELATED
Pleasant Valley police fficer shot...
Packs of stray dogs roam ansas City...
Dec. 12, 2014 Editorial: I-70 il...
Missouri man relieved, verjoyed...
Extremely matted dog und in park...

Click here to watch the story

Iraq War Veteran Aaron Schneider saw a dog get his on I-470 earlier this week, and darted across several lanes of traffic to save it. The oncoming traffic couldn't stop Schneider from saving the beagle.

Schneider said he called animal control, but no one showed up even after 3 hours. So he decided to take matters in his own hands.

Schneider placed the injured dog onto a makeshift stretcher in the back of a truck and took the beagle to get medical care.

he dog suffered from internal bleeding, several broken ribs and a torn windpipe. It remains in itensive care, but thanks to Schneider's quick thinking, the dog is expected to make a full recovery

'rom when I saw him get hit, my heart went out to him," Schneider said. That's why I was so xoused on him. I didn't even think about traffic. It was busy too. I can remember looking down the iohwav and all these cars were comina at me and honkina and flashina their liahts. swervina out o

Behind the Cover: Martin Schoeller and the Three-Legged ...
proof.nationalgeographic.com/.../behind-the-cover-martin-sc...
2040 × 2720 - Jun 26, 2014 - Picture of the photographer using tennis balls to help Layka the dog pose ... of all dog breeds—graces the June cover of National Geographic ...

Editor's Note: A Soldier's Best Friend – National Geographic ...
www.[illegible].com/editors-note-a-soldiers-best-friend-nat...
2040 × 2720 - Nov 24, 2014 - Editor's Note: A Soldier's Best Friend – National Geographic Magazine ... Layka, the dog pictured on our cover and here, was two when she ...

This Three Legged War Hero Will Win Your Heart ...
www.picturecorrect.com/.../photo-three-legged-war-hero-will...
200 × 200 - Nov 11, 2014 - "The close-up was harder than I thought it was going to be, the dog was so full of energy," Schoeller told National Geographic. "[She was] like ...

This three-legged decorated war hero had one leg lost to ...
imgur.com/gallery/[illegible]
2040 × 2720 - report, permalink. That's a dog who ain't got time for none of that "who's a good boy" shit. + ... Was this the dog on the cover of National Geographic magazine?

For portrait photographer Martin Schoeller, known for his signature eye-to-eye, full-face portraits, the least of his worries is having his subject stand still on the X marked on his seamless background. But when it came to a three-legged decorated war hero, one leg lost to surgery after taking four rounds from an AK-47, this was exactly the problem.

Layka, a Belgian Malinois, a breed known to have inexhaustible energy—the highest of all dog breeds—graces the June cover of National Geographic magazine. "She was a celebrity in her own right," Martin said, and after capturing a few frames, "I knew I had my hands full."

Martin was going for a dignified look, fitting of a war dog that almost gave her life to save a squad of soldiers in Afghanistan. "The dog is basically a soldier and we treated her with the respect of a military person," Martin said. "She's not a lap dog that sits on the sofa, but a disciplined fighting dog." He wanted to capture her in this spirit. To do that, Martin knew he had to photograph her with her mouth closed. And that was the tough part.

Fig. 3 External content from image and URL. The *top* submission has a URL and we extracted the URL content. The *bottom* submission has an image and we extracted the text of the image from Google Reverse Image Search

OSNs, like Twitter or Facebook, allow circulating information through a known cycle (e.g., "follow" connections), whereas Reddit promotes a stream of links to all users in a simple bookmarking interface. This makes Reddit a effective resource to study the spread of memes in OSNs [19, 20]. To the best of our knowledge, no similar work has been done on clustering memes in Reddit.

4 Google Tri-gram Method

Google Tri-gram Method (GTM) [30] is an unsupervised corpus-based approach for computing semantic relatedness between texts. GTM uses the uni-grams and tri-grams of the Google Web 1T N-grams corpus [30] to calculate the relatedness between words, and then extends that to longer texts. The Google Web 1T N-grams corpus contains the frequency count of English word n-grams (unigrams to 5-g) computed over one trillion words from web page texts collected by Google in 2006.

The relatedness between two words is computed by considering the tri-grams that start and end with the given pair of words, normalizing their mean frequency with unigram the frequency of each of the words as well as the most frequent unigram in the corpus as shown in Fig. 4, where $C(\omega)$ is the frequency of the word ω. $\mu_T(\omega_1, \omega_2)$ is the mean frequency of trigrams that either start with ω_1 and end with ω_2, or start with ω_2 and end with ω_1. $\sigma(a_1, \ldots, a_n)$ is the standard deviation of numbers $a_1, \ldots, a_n$, and $C_{\max}$ is the maximum frequency among all unigrams.

$$\mathrm{GTM}(\omega_1,\omega_2) = \begin{cases} \dfrac{\log \frac{\mu_T(\omega_1,\omega_2)C_{\max}^2}{C(\omega_1)C(\omega_2)\min(C(\omega_1)C(\omega_2))}}{-2\times\log\frac{\min(C(\omega_1),C(\omega_2))}{C_{\max}}} & \text{if } \log\dfrac{\mu_T(\omega_1,\omega_2)C_{\max}^2}{C(\omega_1)C(\omega_2)\min(C(\omega_1)C(\omega_2))} > 1 \\ \dfrac{\log 1.01}{-2\times\log\frac{\min(C(\omega_1),C(\omega_2))}{C_{\max}}} & \text{if } \log\dfrac{\mu_T(\omega_1,\omega_2)C_{\max}^2}{C(\omega_1)C(\omega_2)\min(C(\omega_1)C(\omega_2))} \leq 1 \\ 0 & \text{if } \mu_T(\omega_1,\omega_2) = 0 \end{cases}$$

Fig. 4 GTM semantic similarity calculation [30]

GTM computes a score between 0 and 1 to indicate the relatedness between two texts based on the relatedness of their word content. For given texts P and R where $|P| \leq |R|$, first all the matching words are removed, and then a matrix with the remaining words $P' = \{p_1, p_2, \ldots, p_m\}$ and $R' = \{r_1, r_2, \ldots, r_n\}$ is constructed where each entry is a GTM word relatedness $a_{ij} \leftarrow \mathrm{GTM}(p_i, r_j)$.

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

From each row $M_i = \{a_{i1} \cdots a_{in}\}$ in the matrix, significant elements are selected if their similarity is higher than the mean and standard deviation of words in that row:

$$A_i = \{a_{ij} | a_{ij} > \mu(M_i) + \sigma(M_i)\},$$

where $\mu(M_i)$ and $\sigma(M_i)$ are the mean and standard deviation of row i. Then the document relatedness can be computed using:

$$\mathrm{Rel}(P,R) = \frac{(\delta + \sum^m a_{i=1}\sigma(A_i)) \times (m+n)}{2mn}$$

where $\sum^m a_{i=1}\sigma(A_i)$ is the sum of the means of all the rows, and δ is the number of removed words when generating P' or R'.

5 Semantic Jaccard Coefficient

Jaccard similarity coefficient is a statistic used to compute the similarity and diversity between two sets. Chierichetti et al. [15] showed that finding an optimal solution for weighted Jaccard median is an NP-hard problem and presented a heuristic algorithm to speed up the computational complexity. The Jaccard coefficient between two sets A and B is defined as follows:

$$J(A,B) = \frac{A \cap B}{A \cup B} \text{ where } 0 \leq J(A,B) \leq 1$$

We propose a modified version of Jaccard coefficient that exploits the use of semantic similarity using GTM. As the original Jaccard coefficient only uses an exact pattern matching, it does not work well if two Wikipedia concepts are not the same but are semantically similar. For example, the Jaccard coefficient for two concepts "President of the United States" and "Barack Obama" should be high as they are semantically similar using GTM.

For two submissions $S_1 = \{T_{11}, T_{12}, \ldots, T_{1n}\}$ and $S_2 = \{T_{21}, T_{22}, \ldots, T_{2n}\}$ where T_i is a Wikipedia concept extracted from the title or comments of submission S_i, the Semantic Jaccard Coefficient (SJC) is defined as:

$$\mathrm{SJC}(S_1,S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2} \text{ where } 0 \leq \mathrm{SJC}(S_1,S_2) \leq 1 \tag{1}$$

where T_{1i} and T_{2j} are semantically equivalent, $T_{1i} \equiv T_{2j}$, if $\mathrm{GTM}(T_{1i}, T_{2j}) \geq e$, where e is a parameter that is explored through the experiment. If T_i is semantically similar to more than one concept in S_2, we use the concept with the highest GTM score.

6 Similarity Scores and Combination Strategies

This section explores the use of GTM semantic similarity of Dang et al. [18] and introduces Wikipedia concepts as an external knowledge to propose five semantic similarity scores and their combinations between submissions. Representing a submission S in Reddit as a vector $S = (T, M, I, U, W)$ where:

- T is an n-dimensional feature vector $t_1, t_2, \ldots t_n$ representing the title of the submission and its description.
- M is an n-dimensional feature vector $m_1, m_2, \ldots m_n$ representing the comments of a submission.
- U is an optional n-dimensional feature vector $u_1, u_n, \ldots u_n$ representing the external URL content of a submission.
- I is an optional n-dimensional feature vector $i_1, i_2, \ldots i_n$ representing the image content of a submission. This content is extracted by using Google reverse image search, which takes an image as a query and extracts the text content of the website that is returned from the top search result and is not from Reddit.
- W is an optional n-dimensional feature vector $w_1, w_2, \ldots w_n$ representing the Wikipedia concepts of the titles and comments of a submission.

6.1 Similarity Scores

We propose five similarity measures between two submissions S_1 and S_2:

- **Title similarity** SC_t is the GTM semantic similarity score between the title word vectors T_1 and T_2.
- **Comment similarity** SC_m is the GTM semantic similarity score between the comment word vectors M_1 and M_2.
- **URL similarity** SC_u is the GTM semantic similarity score between the URL content word vectors U_1 and U_2.
- **Image similarity** SC_i is the GTM semantic similarity score between the word vectors I_1 and I_2 retrieved from Google Reverse Image Search.
- **Wikipedia similarity** SC_w is the SJC score between the bag of concept vectors W_1 and W_2 retrieved from titles and comments of submissions using Eq. (1).

6.2 Combination Strategies

The main goal of this section is to study the effect of different similarity scores and their combinations on the quality of the meme clustering tasks. We incorporate Wikipedia concepts as an external knowledge to all combination strategies from our previous work [18].

6.2.1 Pairwise Maximization Strategy

The pairwise maximization strategy chooses the highest among the title, comment, URL, and image scores to decide the similarity between two submissions. This strategy avoids the situation where similarity scores have a low content quality (e.g., titles are short and lack details, comments are noisy, images and URLs are not always available) by choosing the most similar among them.

Given two submissions $S_1 = \{T_1, M_1, I_1, U_1, W_1\}$ and $S_2 = \{T_2, M_2, I_2, U_2, W_2\}$, the pairwise maximization strategy between them is defined as:

$$\mathrm{MAX}_{S_1S_2} = \mathrm{MAX}(\mathrm{GTM}_{T_1T_2}, \mathrm{GTM}_{M_1M_2}, \mathrm{GTM}_{U_1U_2}, \mathrm{GTM}_{I_1I_2}, \mathrm{SJC}_{W_1W_2}) \quad (2)$$

where $\mathrm{GTM}_{T_1T_2}, \mathrm{GTM}_{M_1M_2}, \mathrm{GTM}_{U_1U_2}, \mathrm{GTM}_{I_1I_2}$ are the title, comment, URL, and image similarity scores between the two submissions S_1 and S_2. $\mathrm{SJC}_{W_1W_2}$ is the SJC score between two submission S_1 and S_2 using Eq. (1) for the Wikipedia concepts extracted from submission titles and comments.

6.2.2 Pairwise Average Strategy

The pairwise average strategy computes the average value of the five pairwise similarity scores. This strategy balances the scores among the five similarities in case some scores do not reflect the true content of the submission. It is defined as follows:

$$\mathrm{AVG}_{S_1S_2} = \mathrm{AVG}(\mathrm{GTM}_{T_1T_2}, \mathrm{GTM}_{M_1M_2}, \mathrm{GTM}_{U_1U_2}, \mathrm{GTM}_{I_1I_2}, \mathrm{SJC}_{W_1W_2}) \quad (3)$$

6.2.3 Linear Combination Strategy

In the linear combination strategy, users can assign different weighting values manually. For example, if users think the title text does not capture the topic of a submission, they can assign a low weight factor (e.g., 0.1). If they think comment texts are longer and represent the topic better, they can assign a higher weight factor (e.g., 0.6). The linear combination strategy is defined as follows:

$$\mathrm{LINEAR}_{S_1S_2} = \mathrm{LINEAR}(w_t\mathrm{GTM}_{T_1T_2}, w_m\mathrm{GTM}_{M_1M_2}, w_u\mathrm{GTM}_{U_1U_2}, w_i\mathrm{GTM}_{I_1I_2}, w_w\mathrm{SJC}_{W_1W_2}) \quad (4)$$

where w_t, w_m, w_u, w_i, and w_w are the weighting factors for titles, comments, images, urls, and Wikipedia concepts with a normalization constraint $w_t + w_m + w_u + w_i + w_w = 1$.

6.2.4 Internal Centrality-Based Weighting

Computing the optimized weight factors for the linear combination strategy is an intensive task. JafariAsbagh et al. [31] used a greedy optimization algorithm to compute the optimized linear combination for the task of clustering memes. However, it is unrealistic to compute all the possible weighting combinations for Eq. (4). To alleviate this computational cost, we propose the Internal Centrality-Based Weighting (ICW), a novel strategy to automatically calculate the weight factors of the linear combination strategy. This strategy calculates the weight factors for each element of a submission by considering its surrounding context. Although all elements of a submission are semantically related, some elements could have more semantic content than others; for example, the URL content discusses more the topic than the title. More weight is assigned to the elements with higher semantic content. The proposed strategy is shown in Eq. (5). It computes the semantic content weights using internal and external similarity scores between titles, comments, URLs, images, and Wikipedia concepts of two submissions. We append all the Wikipedia concepts together to compute the GTM score between Wikipedia concepts and other texts. For each submission, this strategy computes the centrality score for each element of each submission S_i:

$$
\begin{aligned}
\mathrm{CENT}_{T_i} &= \mathrm{GTM}_{TM_i} + \mathrm{GTM}_{TU_i} + \mathrm{GTM}_{TI_i} + \mathrm{GTM}_{TW_i} \\
\mathrm{CENT}_{M_i} &= \mathrm{GTM}_{MT_i} + \mathrm{GTM}_{MU_i} + \mathrm{GTM}_{MI_i} + \mathrm{GTM}_{MW_i} \\
\mathrm{CENT}_{U_i} &= \mathrm{GTM}_{UT_i} + \mathrm{GTM}_{UM_i} + \mathrm{GTM}_{UI_i} + \mathrm{GTM}_{UW_i} \\
\mathrm{CENT}_{I_i} &= \mathrm{GTM}_{IT_i} + \mathrm{GTM}_{IM_i} + \mathrm{GTM}_{IU_i} + \mathrm{GTM}_{IW_i} \\
\mathrm{CENT}_{I_w} &= \mathrm{GTM}_{WT_i} + \mathrm{GTM}_{WM_i} + \mathrm{GTM}_{WU_i} + \mathrm{GTM}_{WI_i}
\end{aligned}
$$

Then, it computes the weighting factors between two submissions S_1 and S_2 by:

$$
\begin{aligned}
w_T &= \mathrm{CENT}_{T_1} * \mathrm{CENT}_{T_2} \\
w_M &= \mathrm{CENT}_{M_1} * \mathrm{CENT}_{M_2} \\
w_U &= \mathrm{CENT}_{U_1} * \mathrm{CENT}_{U_2} \\
w_I &= \mathrm{CENT}_{I_1} * \mathrm{CENT}_{I_2} \\
w_W &= \mathrm{CENT}_{W_1} * \mathrm{CENT}_{W_2}
\end{aligned}
$$

Then, it normalizes the weighting factors so that: $w_T + w_M + w_U + w_I + w_W = 1$, and finally computes the ICW strategy:

$$
\mathrm{ICW}_{S_1S_2} = \mathrm{ICW}(w_T\mathrm{GTM}_{T_1T_2}, w_M\mathrm{GTM}_{M_1M_2}, w_U\mathrm{GTM}_{U_1U_2}, w_I\mathrm{GTM}_{I_1I_2}, w_W\mathrm{GTM}_{W_1W_2}) \quad (5)
$$

6.2.5 Similarity Score Reweighting with Relevance User Feedback

One effective way to improve the clustering results is to manually specify the relationships between pairwise documents (e.g., must-link and cannot-link) to guide the document clustering process [7]. As social network data are intrinsically heterogeneous and multidimensional, it is not easy to compare two submissions to determine if they are similar or not without putting them into the same context. To overcome this limitation, a novel technique, the Similarity Score Reweighting with Relevance User Feedback (SSR), is proposed to incorporate relevance user feedback by a visualization prototype in which submissions are displayed as a force-directed layout graph where:

- **A node** is a submission in Reddit.
- **An edge** is a connection between two submissions if their similarity scores are above a threshold (default 0.85).
- **A node color** represents to which cluster it belongs.

Algorithm 1 describes how the visualization system integrates user feedback to remove outliers, move submissions from a cluster to another, or reassign similarity score weighting factors for submissions. Users can select any of the five proposed

Algorithm 1 Semi-supervised similarity score reweighting with relevance user feedback strategy (SSR)

```
Input: a set of submissions X from Reddit.
Output: K clusters $\{X\}_{l=1}^{K}$
loop
    {Step 1} Perform k-means clustering on P percent of the ground-truth dataset using one of the proposed strategies. P is defined through experiments.
    {Step 2} Visualize the clustering result in step 1.
    {Step 3} Allow users to interactively remove outlier submissions, reassign submission class labels, or assign weight factors for each element between two submissions.
    {Step 4} Re-cluster the submissions based on user inputs.
    {Step 5} repeat step 1 if necessary.
end loop
{Step 6} Recluster the whole dataset considering user feedback in Step 1 to 5.
```

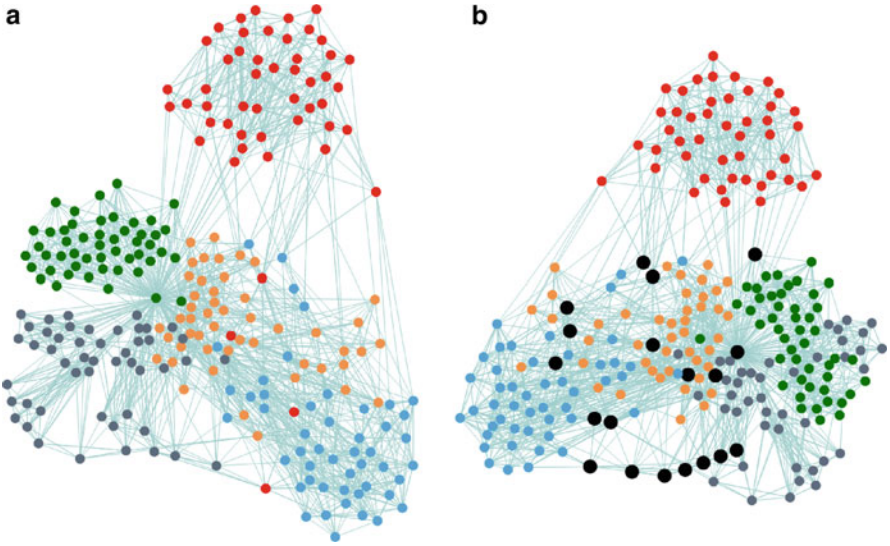


Fig. 5 The proposed meme visualization: (**a**) The original visualization graph and (**b**) The clustering result using ICW

strategies, MAX, AVG, LINEAR, and ICW as a baseline for clustering. Figure 5a shows an SSR visualization of the meme dataset using ICW strategy. The graph has five different colors that represent five memes in the ground-truth dataset. Users can pan, zoom, or click on a submission to get more details about this submission. They can also click on the checkbox "Show wrong cluster assignments" to see which submissions are incorrectly assigned by the ICW strategy. Based on the graph visualization, users can understand how a submission is positioned regarding its neighbor submissions. When clicking on a node in the graph, users will be redirected to the actual submission in Reddit to find out more information and decide if it belongs to the correct cluster. Most of the incorrectly clustered are overlapped or outlier nodes as shown in Fig. 5b. For each incorrectly assigned submission, users can remove, update its class label, or assign a different similarity coefficient score for each element between two submissions. SSR focuses on human knowledge to detect outliers or borderline submissions.

7 The Offline–Online Meme Detection Framework

The meme detection problem is defined for any social media platform used to spread information. In these systems, users can post a discussion or discuss a current submission. An overview of the proposed meme detection framework is shown in Fig. 6.

7.1 *The Offline–Online Meme Clustering Algorithm*

As OSN data is changing and updating frequently, we modify and extend the proposed ICW algorithm [18] to work with the online streaming clustering algorithm using the semantic similarity and Wikipedia concepts to handle continuously evolving data over time. As emerging events or topics are changing in real time, some topics may appear but not burst. Other topics or events may appear and become a popular topic for a long period. Based on this observation, the proposed framework adopts the damped window model [23] and assigns more weight to recent data and popular topics. It also adopts the Offline–Online components of Aggarwal et al. [2] to make the online meme clustering more efficient. As clustering OSN data is a computationally intensive task, the offline component does a one-pass clustering for existing OSN data in the first step. It also calculates and summarizes each cluster statistics using Wikipedia concepts extracted from the titles and comments of all

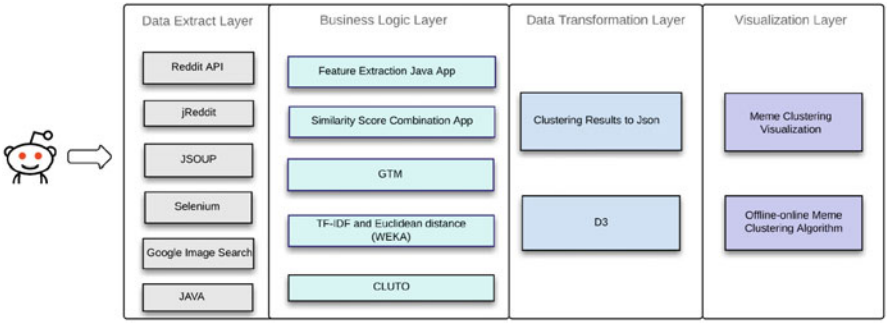

Data Extract Layer
Reddit API
jReddit
JSOUP
Selenium
Google Image Search
JAVA
Business Logic Layer
Feature Extraction Java App
Similarity Score Combination App
GTM
TF-IDF and Euclidean distance (WEKA)
CLUTO
Data Transformation Layer
Clustering Results to Json
D3
Visualization Layer
Meme Clustering Visualization
Offline-online Meme Clustering Algorithm

Fig. 6 The proposed meme detection framework