Mining Representative Posts in Sina Weibo

Yong REN†, Nobuhiro KAJI†, Naoki YOSHINAGA†, and Masaru KITSUREGAWA†

† Institute of Industrial Science, the University of Tokyo
4–6–1 Komaba, Meguro–ku, Tokyo 153–8505 Japan
E-mail: †{renyong,kaji,ynaga,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract Sina Weibo is a popular microblogging service. A great number of people are making use of it to express their opinions, which makes Sina Weibo an extremely valuable source for learning thoughts of crowds on given topics. One chanllendge is finding the important ones form a large amount of Weibo posts. In this work we investigate approaches on selecting representative posts. Several methods including LexRank, Turn down the Noise (TDN) and Kmeans are evaluated. We believe our work is helpful for research such as opinion summarization and information retrieval.

Key words Representative Posts Mining, Sina Weibo

1. Introduction

Sina Weibo(注1) is a prevalent Microbloging service in China. Though it was launched by SINA Corporation on 14 August 2009, it has more than 250 million registered users as of October 2011 according to (注2). More and more people are sharing their opinions on many kinds of topics, which makes Sina Weibo become an important tunnel to collect and learn viewpoints of people on given topics.

But it is an challenging task. One cause is for a given hot topic, usually there are a large amount of posts. And people will express their opinions through different viewpoints. It is both labor and time consuming even impossible to skim them one by one. This brings the well-known problem of information overload. For instance, when we searched "facebook 上市" (注3) using Sina Weibo Search, we could found there are 192,650 related posts as for Feb 17th 2012.

In order to solve the information overload problem, we could make use of ranking function provided by Sina Weibo Search. Users could choose to present related related posts according to hot posts standard (a facility provide by Sina Weibo Search), but there are still 50 pages results (one page contains 20 posts). Further more, for the detail of ranking standard is unknown to outsides, it is still in question if this function could really meet the needs of users. The other choice is only reading posts from famous persons or people that are thought to have enough experience in concerning domain. But unfortunately for kinds of reasons, this approach could not guarantee that users receive un-biased viewpoints. Moreover, such professionals may not appear in Sina Weibo or have not express their opinions.

In this paper, we investigate the approaches for finding the representative post content. Even people discuss a given topic from various viewpints, there is still similarity among those discussions; in other words the content of opinions from different people will have a certain degree of relatedness. Moreover not all those opinions have equal status, some of them should be more comprehensive and representative than others. Select the representative post and identify the number of similar ones could meet the ordinary requirement of both individuals and organizations. And we believe our study will be helpful for research such as opinion mining, information diffusion, and recommendation reading.

At first glance, text summarization that can catch the main points in documents would be a proper choice. We exploit one text summarization called LexRank [3] in our study. Conventional clustering algorithm such as Kmeans could also be one potential choice. We could firstly cluster Weibo posts, then choose posts from main clusters as representative ones. Another option is choosing several posts that cover related concepts as much as possible, related methods usually involve optimization problem. We will further discuss these approaches in section 4.

One potential application of our work is in special report which is usually involved in collecting, editing and presenting important and useful content to viewers. One real instance is present in figure 1, it is the special report of Sina for facebook

(注1)：*http : //www.Weibo.com/*
(注2)：*http : //www.ciccorporate.com/*
(注3)：In Chinese language 上市 means IPO

IPO. (注4). Important and useful Weibo posts (The title for them here is called hot discussion in Sina Weibo) are listed in the "special report" together with other forms of content such as news report from main websits (For the space is limited, we do not show the whole picture, you can find the while special report in the link listed in footnote below). We should point out that those content selecting process are done by professionals. We hope they could benefit from our work.

The rest of this paper is organized as follows: Section 2 introduces related work,Section 3 give the preliminaries involved in our work. Section 4 explains the approaches we adopt in detail. Section 5 shows the experiment result and discussions. Section 6 concludes this study and discusses future direction.

2. Related Work

There are several studies whose objectives are similar to our work.

The authors in [3] proposed an unsupervised summarization approaches called LexRank. The main idea of LexRank is to find important sentences in a given document to take them as summarization of that document. Firstly they treate each sentence as bag-of-words with TF-IDF weighting and computed cosine similarity score between any two sentences in that document. A similarity graph among sentences is built based on the similarity score. Then they exploited a PageRank-like algorithm to rank sentences according to rank value of each sentence. The difficulty in applying LexRank to our task lies in the length of Weibo post is usually short, and the language usage is diverse and irregular. In order to conquer these issues, we tried LDA to get additional features. Extracting appropriate features from short Weibo posts and compute the similarity among posts is still challenging in our study.

The authors in [2] described the algorithm of Turn Down the Noise (hereafter we use TDN to referrer to Turn Down the Noise as the authors do). The purpose of this work is to resolve information overload brought by large scale of blogsphere. They proposed a notion of coverage and formalized it as a submodular optimization problem. We will further explain the concept of submodularity in section 3. Their goal is to show users a small set of blog posts covering hot topics. Though the purpose seems quite similar, content coverage and content representativity are not the same, which can be seen in experiment results in section 5.

We evaluated LexRank, TDN and Kmeans one a real dataset. Representative posts selected by using these algorithms will be compared. Though these approaches have been put into usage in works that have similar purpose with our task. It is still in question whether they could perform well in the situation where the text is brief, limited contexts are provided and words usage are various.

(注4)：*http : //tech.sina.com.cn/z/facebook_ipo/*

Fig. 1 Special report of Sina about Facebook IPO

Number	Representative Post
1	【Facebook 正式提交 IPO 申请 拟融资 50 亿美元】Facebook 上市将是美国历史上最大规模的科技公司 IPO 交易，市场预计估值可能会达到 750 亿-1000 亿美元。2004 年谷歌上市融资 19 亿美元，而当时市值为 230 亿美元。摩根士丹利、摩根大通、高盛、美银美林、巴克莱资本以及 Allen 将担任 Facebook 上市交易的股票承销
2	【#Facebook 上市#】Facebook 今日已向美国证券交易委员会(SEC)正式提交 IPO 申请，计划融资 50 亿美元，股票代码 FB，上市交易所暂未确定。Facebook 上市将是美国历史上最大规模的科技公司 IPO 交易，Facebook 估值可能会达到 750 亿-1000 亿美元。
3	#Facebook 上市#我佩服他们的创意，一个简单的社交网站居然能达到如此高峰。南京现在也在搞创业，但是仅仅通过一些激励貌似还有点力不从心，改变一下学校里授课方式，或许会让更多人接触到创业这么一个事物。
4	#Facebook 上市#呵呵，不知道这么多人跟着瞎起哄干什么,真正用过和了解facebook 的又有多少，又是抱着太监看隔壁做爱的心态吧
5	涂鸦艺术家 David Choe2005 年给 Facebook 总部画墙画，工钱为几千美元。但他没有要现金，而是要了当时等价的 Facebook 内部股份。Facebook 上市后，他的股票的市值估计会达到 2 亿(200 million)美元。见 http://t.cn/zOhJNou
6	Facebook 创下全球互联网公司上市融资纪录；上市估值约为 750 亿至 1000 亿美元。Facebook 上市不仅会创下科技公司上市之最，也将造就至多千名百万富翁。根据美国证交会的一份文件，扎克伯格拥有 28.4%的股权，若公司估值达到最高上限，他个人股票价值将高达 284 亿美元，即刻跻身福布斯前十名。
7	【Facebook 上市】一直以来关于这家全球最大社交网站 IPO 的消息不断，最新的情况是：Facebook 将于 2 月 1 日提交上市申请，该公司估值为 750 亿美元至 1000 亿美元，摩根士丹利或成为其主承销商，但高盛也将在交易中扮演“重要角色”。http://t.cn/z0sr9Xg
8	【Facebook 上市在即 李嘉诚有望大赚 5 倍】知名社交网站 Facebook 可能最早将于本周三提交首次公开招股(IPO)申请，而上市后的估值最高将达 1000 亿美元。香港首富李嘉诚曾通过私人基金于 2007 至 2008 年以 1.2 亿美元投资 Facebook，若公司成功上市，李嘉诚有望在 Facebook 上大赚超过 5 倍。http://t.cn/zOPG2T8
9	#Facebook 上市#Facebook 招股书里列的四个限制访问国家是中国，伊朗，朝鲜和叙利亚。这说明啥？
10	Facebook IPO 是美国历史上规模最大的公开发行之一，为了成为此次 IPO 的主承销商，华尔街各大投行进行了激烈竞争。最后摩根士丹利出任 Facebook 上市交易的主承销商，而摩根大通、高盛、美银美林、巴克莱资本以及 Allen&co。略微有些惊讶的是，佣金比例仅仅为 1%。 #i 读报#

Table 1 Representative post from LexRank

3. Preliminaries

Given one topic T and related posts $P = \{p_1, p_2, ...p_n\}$, where p_i is one post. Our purpose is to find a set of posts $R \subset P$ that can catch the main themes in the set of posts P. And we should emphasize that the size of R should be much smaller than the size of P.

In Sina Weibo users are using diverse vocabulary to describe things. When we compute content similarity between

Number	Representative Post
1	【聚焦每日热点，@校导网 带你读新闻】2 月 3 日热点：facebook 上市,冲击千亿美元市值、小肥羊退市，百姓无奈惋惜、全国血荒蔓延，伤透民心、张默吸毒，依法拘留 13 天、响水惊现多具儿童尸体、继宰客门，三亚再爆天价菜单、北大《女生日记》网络走红、梁朝伟生活体曝光……全文地址：http://t.cn/zOh63zY
2	Facebook 上市之 15。在我看来，Facebook 等不少成功的硅谷互联网公司都有一种共同的理想主义的气质，那就是我只管朝改变世界的远大目标前进，不用拼爹、拉关系、讨好旧势力，站着也把钱挣了。社会化网络有力地打击了历史悠久的信息不对称而造成的权利不公平问题，的确有可能使人类社会往好处去。
3	分享图片在剩下的一个学期四个月里，你愿意与学习结为伴侣，无论贫穷还是富贵，无论电脑还是手机，无论多困或者多累，无论想吃还是想睡，都要把学习放在第一位，以不挂科为目标，同甘共苦同舟共济永不言弃，爱惜她尊重她理解她保护她，你愿意这样做么？Yes,I do! #facebook 上市# #张默涉嫌吸毒#
4	小学时候体育课，一同学买了包袋装酸奶，没来及喝，上课铃响了。他当时也不知怎么想的，顺手就吧酸奶放进自己戴着的帽子里了！结果课上到一半，他不老实，逗弄女生被体育老师发现。老师冲上来就朝他脑袋拍了一巴掌！顿时酸奶就顺着脑袋彪下来了！把体育老师吓的大叫着退了三大步#德纲体##Facebook 上市#
5	#Facebook 上市#赢战 2012，家居建材行业风云再起，下一个霸主将会是谁！网络营销策划人赵武林依托赢道顾问，专注企业品牌建设，力推“品牌加速度计划”，助您成就霸业！目前我公司已相继服务金舵陶瓷、皇朝家私、3A 环保漆等数十家企业。成功热线：15228977723；QQ954870381；或直接回复
6	想过普通的生活，就会遇到普通的挫折。你想过上最好的生活，就一定会遇上最强的伤害。这世界很公平，你想要最好，就一定会给你最痛。能闯过去，你就是赢家，闯不过去，那就乖乖退回去做个普通人吧。所谓成功，并不是看你有多聪明，也不是要你出卖自己，而是看你能否笑着渡过难关。早安~#Facebook 上市#
7	Facebook 上市在即，一些早期投资者将获得上千倍的回报，但最独具慧眼，最善于“价值”发现的人，其实是扎克伯格的华裔女友普丽西拉·陈，她得到了无价之宝。两人同为哈佛学生，在派对厕所派对时相识。陈的外貌并不突出，却成功抓住这位理科男生硅谷极客的心，并对他的事业鼎力相助，堪称一段佳话。
8	#深度阅读#热闹了一天，静下心来看看#Facebook 上市#这点事儿。有人因它一夜暴富，有人为此后悔不迭；有人说它估值过高，也有人说好戏刚开演。附 Facebook 提交上市申请概况、招股书译文、扎克伯格的长信，点进微博查看各大名刊精简版评论长微博，重温盛宴或为明天留一点谈资。http://t.cn/zOhfKHK
9	#Facebook 上市#我叫白云，我叫黑土，谁知盘中餐，粒粒皆辛苦。会当凌绝顶，大城市铁岭。十年生死两茫茫，不思量，猪撞树上，你撞猪上！床前明月光，我叫不紧张。南朝四百八十寺，此处省略一万字。多行不义必自毙，一场大水没咋地。小楼昨夜又东风，他把猴屁股当红灯
10	#Facebook 上市#FB 团队，用 8 年时间，通过 使用区域拓展、设备升级、技术创新、赢得连续风投、抢得微软合作、植入广告盈利模式、收购创新小企业、推出虚拟货币、拓展手机应用平台、参与时政……等手段，个人感叹：莫欺少年穷

Table 2 Representative post from greedy algorithm

Number	Representative Post
1	#Facebook 上市# 马克·扎克伯格，美国社交网站 Facebook 的创办人，被人们冠以“盖茨第二”的美誉。哈佛大学计算机和心理学专业辍学生。目前 Facebook 已启动上市进程，若公司估值达到最高上限，他个人股票价值将高达 284 亿美元，是全球最年轻的单身巨富，也是历来全球最年轻的自行创业亿万富豪。留做资料吧
2	#容斋喜乐会#【Facebook 上市造富奇迹：漆墙工获股票价值 2 亿美元】2005 年,一位叫大卫•乔伊的员工受 Facebook 的邀请,前往公司在加州帕洛艾洛的总部进行墙面装饰.作为报酬,公司向乔伊提供价值数千美元的股票,尽管乔伊当时认为 Facebook 荒谬且没有意义,但仍然收下了股票,如今这些股票已价值 2 亿美元。。。
3	#Facebook 上市#人家都上市估值千亿了，可是大陆还是封了 Facebook，所以我们聊什么都没有什么意义，只能用新浪微博还不是和 Facebook 差不多的，呵呵 0(∩_∩)0~
4	#Facebook 上市#之如何胜出：①产品有用、易用、性能和稳定性。②用户数和组成。③用户粘性。④产品时机和市场接受度。⑤产品尤其是移动的赢利能力。⑥广告频度、尺寸和显著性。⑦客服。⑧市场与销售。⑨开发者关系。⑩兼并。⑪吸引优秀员工尤其是软件工程师。⑫低成本运营。⑬品牌与声誉。(转)
5	FACEBOOK 上市之后，六千员工都将一夜暴富。最离奇的是：当初给总部办公室刷壁画的那位，明智地选择要股票，而不是现金。七年前的空头承诺现在价值两亿美金!! 反面典型是创始人在大学宿舍的室友，让他辍学帮忙经营他不肯，错过了几十亿美元。看着他接受记者采访，真想问：你怎么还不一头撞死啊？
6	#Facebook 上市#扎克伯格绝对是一个传奇，2010 年的电影《社交网络》便描述了他的创业历程，没看过的同学推荐看看。顺带提一下，这家超过八亿用户、世界上最成功的网站只有四个国家的公民被禁止进入，朝鲜、古巴、伊朗和中国大陆。
7	@吕氏讲堂：Facebook 最大的成功之处，就是把人际之间的分享关系变成了未来人类信息传播的基础设施。——吕本富微访谈到此结束，感谢大家！http://t.cn/zOhZ31a 接下来@胡延平 将做客微访谈继续探讨 Facebook 上市影响，由于当地网络问题略有延迟，请大家谅解，稍等片刻！http://t.cn/zOhAPkg
8	#Facebook 上市#我孤陋寡闻以为脸书早 IPO 了...毕竟人人网去年也在纽交所上市了不是。员工高管的原始股和期权，这次肯定赚发了~！当初的风投者更乐。但话说回来，中国根本上不了 Facebook，没什么好兴奋热议的，瞧瞧热闹就好。
9	#Facebook 上市#大家在谈论这个国内显示是这个 Internet Explorer cannot display the webpage 的网站吗？是很重大的国际新闻，可是对于一个 90%以上的中国人这个上都上不去的网站，讨论这个有什么意义。facebook 就算消亡或者上市了，对于连界面见都没有见过的人谈这个有什么意义。不觉得应该反省吗？
10	#Facebook 上市# 我一直在回忆自己最初的梦想是什么。这么多年，梦想换了很多个，反倒是最初的梦想不记得是什么了

Table 3 Representative post from k-means clustering

two post using methods that rely on common words, words mismatch usually occurs. One example is in Chinese both "電腦" and "計算機" are used to describe computer, but we can not discover the similarity between two posts mentioned "電腦" and "計算機" respectively through similarity measures that rely on common words. In order to solve the words mismatch, we exploit LDA in the similarity computing phase. In LDA, a topic is recognized as probability distribution over words and each document is taken as a mixture over topics. It is easy to find the probability distribution among potential topics for a given text, and the probability value can form feature vector.

A set fuction F is submodular $if, \forall A \subseteqq B \subseteqq V, \forall s \in V \setminus B, F(A \cup \{s\}) - F(A) \geqq F(B \cup \{s\}) - F(B)$. We can take set A ,set B as set of posts, and take s as one invidual single post, as pointed out by the authors "submodularity characterizes the notion that reading a post s after reading a small set of posts A provides more information than reading s after having already read the larger set B." In other words, the concept of submodularity can meet the requirement of choosing a few posts but cover as much information as possible.

Maximizing submodular functions in general is NP-hard [4]. But the work [5] verified that the greedy algorithm could give a approximation solution,we adopted this strategy in our study. More efficient algorithms will be explore in our future work.

4. Approach

We exploited LexRank, Turn down the noise (TDN) and conventional clustering method (Kmeans) in this study. Our method consists of three phases:

- **Text representation**: Since most of posts are short, we do not segment them. We treat them as bag-of-words.
- **Measure similarity**: A similarity matrix is required in LexRank. We use cosine measure to calculate the similarity score.
- **Select representative post**: When we adopted LexRank we equalizes representatively as significance of related post. When we use TDN algorithm, we treat representatively as the concept of coverage defined in this work [2].

LexRank belongs to the area of extractive summarization [1] which involves in identifying central sentences in one document. Those central sentences are believed to give the necessary and sufficient information related to the main idea of one document, this is consistent with our expectation.

We expanded the idea of LexRank by taking each post as one sentence, though some post are composed of multiple sentences. But we should point out that many posts are short and only have one sentence for the limitation of 140 Chinese characters.

In LexRank a similarity matrix is required. We computed similarity matrix using TF-IDF feature vector and LDA feature vector respectively, then we add these two similarity matrixes together. At present we set the weight parameter to 0.5 for both similarity. Through this process, we can catch both word similarity and topic similarity; at the same time it could resolve the words mismatch problem in Weibo posts.

TDN is proposed in the work [2] as one approach to solve optimal coverage problem in blogsphere. The coverage function is defined as:

$$cover_A(i) = 1 - \prod(1 - cover_j(i)) \ where \ a_j \in A$$

$cover_j(i)$ is defined as probability of a feature i given a post i

function F on set A is defined as:

$F(A) = \sum w_i cover_A(i)$

where w_i is used to height the importance of feature i.

According to [2] the fuction F is submodular. One character of coverage defined in their work is it encourages to cover new information.

According to the authors there are two options to represent text when adopt TDN, they called them "high-level features" (topic related feature) and "low-level feature" (words level feature) respectively. Both of them are required to be probability value from generated model. Appropriate "high-level feature" can be easily attainted using LDA, but for the "low-level feature" there are still several issues as how to decide the probability value in our task. So we choose topic related features attained from LDA.

As the space is limited, we skip the detail of Kmeans, it is one well-adopted clustering algorithm. Now we only take LDA feature into account when we did clustering on posts. The reason is we find the TF-IDF feature vector is highly sparse, it is difficult to get meaningful clusters by using Kmeans.

5. Experiment and Discussion

5.1 Data Preparation and Pre-processing

From Feb 2nd 2012, related news about Facebook IPO has attracted much attention from Sina Weibo users, it became one of hottest topics listed in Sina Weibo around Feb 3rd 2012 and Feb 5th 2012. It is that time we collected the posts about the topic "facebook 上市". For there is a large amount posts published in realtime in Sina Weibo, it not an easy way to collect all the posts about Facebook IPO. We collected data through Sina Weibo Search. We issued the query "facebook 上市" into Sina Weibo Search, and ranked the results according to the hotness; then we crawled all the results. We did this several times during heated discussion period.

We deleted duplicated posts through post id. Totally we get 1509 posts. Though the number of posts is small, they come from hottest posts and they are collected during the hot discussion time. We believe we could catch kinds of important and useful ideas from analyzing those posts. In the future, we could collect more related information from these posts, potential approaches include crawl reposts, comments and other recent posts from the users whose posts have entered in the hot posts list.

For those hot posts crawled, we firstly deleted words "facebook"(all the variant such as Facebook are also excluded), "上市" and its English counterpart "IPO". And we further excluded other 30 common words, we found most of them are background words such as "全球"(global), "社交"(social) and so forth. The main reason is those common words will bring disturbance when we compute similarity score between two posts, or cause mis-clustering at the time we choose to group posts. We also excluded shortlinks and "@" content. For the time being, we take all the posts as opinionated sources.

5.2 Experiment Result

The representative posts extracted using different approaches are shown in table 1, table 2 and table 3 respectively. We could find there are no overlapping posts in the three groups of posts. According to our manual evaluation, we found the representivity of posts extracted from LexRank is the best. Possible explanation is LexRank could get the centroid among posts. And the group of posts extracted using greedy algorithm contains the most concepts, in other words content in this group it is the most diverse. It conforms to the purpose of detecting broad information. According to our manual evaluation by now, covering concept as broad as possible is not a recommend choice in finding representative content.

6. Conclusion and Future Direction

In this paper, we introduced the requirement of selecting the representative posts in Sina Weibo. We have investigated the usage of LexRank, TDN and conventional clustering approaches for this task. For the time being, we found LexRank, which is a text summarization approach, could bring the best results.

We also found it is difficult task to build feature vector for Sina Weibo posts. Traditional text analyzing approaches such as LDA do not perform well. Possible causes are firstly Weibo posts are usually short, which means there are not enough contexts that play important role in ordinary text analyzing methods. Additionally, the expressing ways of Sina Weibo users are flexible and various, analyzing methods based on words-level feature are not appropriate.

In the future, we will explore feature vector building methods for short posts. Specifically speaking, given one short post, to some extend, we need to infer the real meaning hidden behind. Outlinks containing in post, repost and comment will be explored, we suppose those enrich information could make up for the context lost because of post length limitation. We will also plan to take semi-supervised learning methods, for example, semi-supervised LDA. One main advantage of semi-supervised learning algorithms is that we could encode some prior knowledge in advance. Those prior knowledge could help us to learn the posts better.

We will further verify our observation at present in large scale dataset. When we have a large amount of data, we will

consider more efficient algorithms, for example, we hope to exploit the Cost-Effective Lazy Forward (CELF) proposed in [5] other than greedy algorithm in using TDN.

In addition, we will take features of users and the relationship between users in Sina Weibo into account. It is reasonable to put more weight on the posts published by users who own much expertise in one domain. It is reasonable to put the opinions of friends into the representative posts list if they have taken part into the discussion of given topic. Because Sina Weibo is a SNS platform, people should be more concerned about the thoughts from friends.

References

[1] D. Das and A. F. T. Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II Course at CMU*, 4:192–195, 2007.

[2] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 289–298, 2009.

[3] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128, 2011.

[4] S. Khuller, A. Moss, and J. S. Naor. The budgeted maximum coverage problem, 1997.

[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 420–429, 2007.