基于数据世系的微博信息管理与检索算法研究

黄庆宇 卢珞先
(武汉理工大学信息工程学院 武汉 430070)

摘 要 在微博平台中用户的消息以流的形式按照时间顺序到达系统,对微博数据流的有效管理可以及时地响应用户的查询操作。基于数据库的数据世系思想,提出了一种基于数据世系的微博信息管理方法。首先,根据事件的产生、发展以及变化,将同一社会事件包含的消息定义为数据世系;其次,将微博消息流划分为不同的数据世系,并根据新消息动态地维护数据世系集合;最后,应用数据世系中的文本消息响应用户的查询。实验表明,基于数据世系的微博信息管理方法使用的内存少,运行效率高,可用于微博消息流的实时处理及查询响应工作。

关键词 数据世系,数据流,微博,信息检索
中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.040

Provenance Based Information Management Method for Microblog Messages

HUANG Qing-yu LU Luo-xian
(School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China)

Abstract In microblog platform, users' messages arrive the system in a temporally ordered sequence, and efficient management of microblog streaming data can handle users' queries timely. Based on provenance of database, a provenance based information management method for microblog messages was proposed. Firstly, the provenance is defined as messages about a common event according to the generation, development and changing of an event. Secondly, the message streaming is divided into different provenances and they are maintained dynamically when a new message comes. Finally, the messages of provenance are used to answer user's queries. The experiments show that the proposed method is efficient in memory usage and time cost, and can be used to timely response of users' queries.

Keywords Provenance, Streaming data, Microblog, Information retrieval

随着智能终端的普及和社会媒体网站的盛行,互联网上出现了大量的微博服务,如 Twitter、新浪微博和 Foursquare 等。在微博服务中,用户可以很容易地发布和接收状态信息,也可以对别人的状态信息(tweets 或者 posts)进行评论或转发。Twitter 仅仅成立几年就拥有了超过 10 亿规模的用户,并且每天用户的状态更新数量为 2.3 亿[1]。这些海量的用户状态更新信息反映了用户的社会活动以及用户的兴趣等。在微博这种社会视角下,研究人员进行了不同方向的分析工作,如社会科学[2]、股票市场预测[3,4]及用户情感分析[5,6]等。

微博的消息是只有 140 字符的短文本,并且用户的消息时序地进入系统。在微博消息流的快速变化中,大量的事件不断涌现,并迅速被新的话题所湮没,因此难免存在噪音和碎片消息。传统的关键字搜索很难应用于微博环境,其原因有两点。首先,微博的消息不超过 140 字符,这使得用户很难理解这些短文本。此外,用户经常用一些词的缩写或者发布无用的消息,由于微博的即时性和公开性,使得平台中含有大量的噪音及低质量消息[7]。其次,微博消息的传播非常容易。微博提供转发服务,这使得消息的传播更为便捷。重大新闻能在很短的时间被众多的用户所浏览。在传播过程中,原始消息的传播以及改变后的消息会产生分离。因此,用户很难有效地理解微博的消息并掌握关于话题的内容。

近期的研究表明,重大事件或者明星的消息在微博中非常流行,用户经常通过反复的搜索行为关注事件或者明星的动向[8],这种搜索方法不同于传统的 Web 搜索。微博服务的各种特性使得微博搜索更具挑战性,其中对这些短文本的有效组织和更直观的解释是最基本的研究内容。本文基于数据世系的相关思想,通过对微博消息的原始及演化信息进行跟踪,提出了一种基于数据世系的微博信息索引方法。

1 相关工作

数据世系描述了数据在整个处理周期内的起源以及发展变化,在数据管理领域中有着广泛的应用。工作流系统采用数据世系来跟踪数据源以及数据变换后的结果。数据世系的概念模型有提取和维护方法[9]。在社会媒体领域,信息流数据比较活跃,并被频繁更新。由于数据的变化复杂以及数据总量大,使得传统的数据世系方法[10]很难应用于微博平台。本文根据微博消息的变化轨迹,设计了相应的摘要索引结构用于发现数据世系。

到稿日期:2014-11-12 返修日期:2014-12-23 本文受国家自然科学基金(61170306),湖北省科技攻关基金项目(2003AA101B05)资助。
黄庆宇 男,主要研究方向为社会网络与信息处理、云计算;**卢珞先**(1962-),女,副教授,主要研究方向为网络与信息处理、网络通信与分布处理技术、云计算服务。

社会媒体数据的大量涌现改变了互联网上信息的创建和使用方法。通过社会媒体，任何用户都可以产生并发布在线网络信息。通过用户的及时发布，任何社会事件都可以迅速地在社会媒体中体现出来。为了分析并理解这些社会事件，突发事件检测和信息级联挖掘已经成为社会网络分析的重要研究内容[1,11]。Leskovec 等人[12]对博客中广泛流行的短语进行调查后发现，与传统的新闻媒体相比，博客覆盖的领域更广，文章的质量更好。Wu 等人[13]提出了一种两阶段的信息传播模型，并在大规模的 Twitter 数据集上进行了验证。用例研究表明，社会媒体中的用户为了了解某一事件的进展，往往对某一事件或人物进行反复搜索[14]。上述的研究大都采用离线的方式对微博数据进行分析和挖掘。为了对事件进行动态的监测及后续的查询操作，本文基于数据世系的理论对消息的产生和发展进行跟踪。

微博消息文本简短并且包含大量的噪音。为了及时判断到来的消息是高质量的消息还是噪音，Chen 等人[15]提出了一种局部索引机制。该方法实时地对高质量的信息建立索引，并通过批处理的方式对噪音进行索引。微博数据管理的另一种方法是应用话题模型[16]来提取微博中的主题概念。消息作者的专业知识、读者的反馈以及消息的广播领域是从微博数据中提取高质量信息的重要因素[17]。然而，这些方法都存在着相应的缺陷，其中动态信息的上下文以及演化过程没有体现出来，基于项的排名方法或者简单的分组方法并不能跟踪消息的传播轨迹。本文应用数据世系索引方法来显式地表达原始数据与演化数据间的联系，通过索引对数据内容进行组织和管理，并用于后续的数据分析和查询支持。

2 基于数据世系的检索方法

微博消息可以按照微博平台的时间轴形成一个大的数据流，并且这些消息之间根据消息的发展及演化存在着一定的联系。本文分别定义了微博消息流、微博消息数据世系以及由相关消息构成的数据世系组。

定义 1 微博消息流是按照发布时间形成的有序的文本序列$\{t_1,\cdots,t_i,\cdots\}$，其中消息 t_i 是一个 6 元组$[date,user,msg,urls,hashtags,rt]$。

定义 2 微博消息数据世系保存着消息之间的发展轨迹。在微博消息流中，新到来的消息与已有的消息通过预先定义的链接标准建立有向的链接，其中 t_i 与 t_j 之间链接的取值可以通过 msg、$urls$、$hashtags$ 或者 rt 得到。

定义 3 数据世系组由一组消息组成，$T=\{t_1,\cdots,t_i,\cdots,t_n\}$，$T$ 中消息之间的链接被保留，而数据世系组之间的链接被忽略。

微博消息流是一个时序的数据系列，这些数据系列元组按照数据世系形成的有向边构成森林。森林中的每一棵树表示一个事件的产生及其发展的整个过程。在微博信息检索过程中，根据用户输入的关键字与每棵树的摘要进行匹配，并将符合用户需求的树返回给用户。

2.1 数据世系构建

在微博消息流中，当一条新的消息到达时，从该消息中提取 msg、$urls$、$hashtags$ 和 rt 等链接指示符以及消息包含的关键字。应用得到的指示符或关键字，将该消息与已知的数据世系进行匹配。如果该消息与某一个数据世系匹配，则将该消息加入到该数据世系中；如果无数据世系与之匹配，那么将该消息作为数据的起源单独构成一个数据世系。

在新消息与数据世系的匹配过程中主要考虑链接的相关性、数据世系的时间跨度以及大小等因素。数据世系的构建主要包含候选数据世系选择、数据世系评分和插入消息 3 个部分。

1)数据世系选择

对于新消息的每一项(包含标识符和关键字)，从数据世系集合中选取包含该项的数据世系并将其加入到候选列表中，于是得到了候选数据世系的列表。给定数据世系集合 I 和新消息 t，数据世系列表的计算如式(1)所示。

$$List(t,I)=\{d\mid d\in I, msg(d)\cap msg(t)\neq\varnothing\vee urls(d)\cap urls(t)\neq\varnothing\vee hashtags(d)\cap hashtags(t)\neq\varnothing\vee rt(d)\cap rt(t)\neq\varnothing\} \tag{1}$$

2)数据世系评分

在候选数据世系中，计算新消息与数据世系的相似性，并根据相似性的大小对候选数据世系进行排名。给定新消息 rt、候选数据世系 B，它们之间的相似性计算公式为

$$S(t,B)=\alpha|date(t)-date(B)|+\beta|msg(t)\cap summary(B)|+\gamma|urls(t)\cap urls(B)|+\lambda|hashtags(t)\cap hashtags(B)|+\omega|rt(t)\cap rt(B)| \tag{2}$$

其中，$date(B)$为 B 中所有消息的平均时间，$summary(B)$包含了 B 中所有消息的 msg，$urls(B)$包含了 B 中所有消息的 url，$hashtags(B)$包含了 B 中所有消息的 $hashtag$，$rt(B)$包含了 B 中所有消息的 rt，参数 $\alpha,\beta,\gamma,\lambda$ 和 ω 为相应因素的权重，并且 $\alpha+\beta+\gamma+\lambda+\omega=1$。在实际的应用中，上述参数可根据实际情况进行人工调整。

3)将新消息插入到数据世系中

由于数据世系中的消息是一个树形的结构，新加入的消息需要加入到树的叶节点。在将新消息加入到树的过程中，需要考虑新消息与所有节点的相似性以及其它相关特征。首先收集数据世系中每个节点消息的匹配信息，然后将该节点的匹配信息与新信息进行相似性计算，最后将该新消息作为与其最相似节点的叶节点加入到树中。

在将 t 加入到 B 的过程中，需要将 t 与 B 中所有的消息 b 进行相似性计算。在计算消息 t 与 b 的相似性时，考虑 msg、$urls$、$hashtags$、rt 和时间 5 种因素，每种因素的相似性分别为

$$S_m(t,b)=\frac{|msg(t)\cap msg(b)|}{|msg(t)|} \tag{3}$$

$$S_u(t,b)=\frac{|urls(t)\cap urls(b)|}{|urls(t)|} \tag{4}$$

$$S_h(t,b)=\frac{|hashtags(t)\cap hashtags(b)|}{|hashtags(t)|} \tag{5}$$

$$S_r(t,b)=\frac{|rt(t)\cap rt(b)|}{|rt(t)|} \tag{6}$$

$$S_t(t,b)=\frac{|1|}{|date(t)-date(b)|+1} \tag{7}$$

得到每个因素的相似性后，通过对不同的因素进行加权，得到 t 与 b 两个消息的整体相似性：

$$S(t,b)=\alpha S_m(t,b)+\beta S_u(t,b)+\gamma S_h(t,b)+\lambda S_r(t,b)+\omega S_t(t,b) \quad (8)$$

2.2 微博消息检索

在构建完数据世系后，微博的流数据形成了若干数据世系(树形结构)组成的集合，并且随着时间的推移这个集合越来越大。然而，通过对大规模微博流数据的分析发现，数据世系的规模(包含的消息个数)具有很大的差异，数据世系的数量与包含的消息个数呈幂率分布。图 1 为数据世系的分布图，横轴为数据世系包含的消息个数(Bundle Size)，纵轴为数据流中数据世系的数量(Provenance Bundle Count)。从图 1 可以看出，大多数的数据世系仅由少量的消息组成。

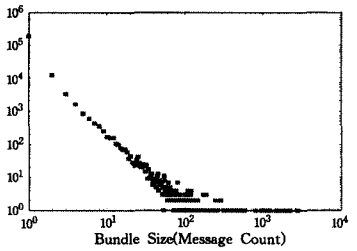


图 1 数据世系的规模分布

数据世系规模的幂率分布表明大多数的数据世系仅由少量的消息组成，仅有部分重大的事件包含数量较多的消息。由于微博消息具有即时性，一个事件在形成并发展结束后很难引起用户的兴趣。利用微博信息的这一特性可以对数据世系的存储和设计进行优化。

微博消息以流的形式源源不断地流入系统，不可能将所有的数据世系都保存在内存中。采用滑动窗口技术保存数据流的一个片段既符合微博事件的即时性又符合数据世系的幂率分布特征。这里的滑动窗口指的是数据世系的滑动窗口，即数据世系池。在内存中保存时间最近的那些数据世系，并随着新消息的加入不断地更新该数据世系池。当数据世系池中的数据世系在较长的一段时间没有更新时，便将其从池中移除并保存在磁盘上。

在应用数据世系响应用户的检索时，将数据世系的摘要信息与用户输入的关键字进行匹配，在匹配时可以应用数据世系的指示符进一步提高检索的性能。数据世系 B 与查询 q 的相关性可通过如下公式计算。

$$r(q,B)=\alpha * s(q,B)+\beta * i(q,B)+(1-\alpha-\beta) * t(B) \quad (9)$$

其中，参数 $0\leqslant\alpha,\beta\leqslant 1$ 且 $\alpha+\beta=1$。

3 实验设计与结果分析

3.1 实验环境与数据集

实验采用的平台为 8 核 CPU 和 32GB 内存的 x64 Linux 服务器。使用的编程语言为 Python 和 Django，微博信息的检索采用 Lucene。

本文采用 Twitter 提供的 API 进行了数据的采集。数据采集的时间为 2009 年 8 月和 9 月，共收集了约 2500 万条微博消息，其中每天的消息大概有 7 万条。

在实验中，微博消息以数据流的形式按照时间序列流入系统，最后到达系统的消息的时间就是系统的当前时间。为了测试数据世系方法的性能，分别实现了如下 3 种算法。

• 全部索引(Full Index)：在数据世系的发现过程中，每次当新消息到达的时候将其插入到最适合的数据世系中。全部索引方法将整个数据集进行离线数据世系发现，对发现的所有数据世系集合进行索引，并用于对查询的相应调整。全部索引方法作为数据世系发现的标准。

• 部分索引(Partial Index)：部分索引方法是流算法。其中内存是有限的。在数据世系的构建过程中，对单个数据世系包含的消息个数无限制，内存中只包含一定数量的近期的数据世系。

• Bundle Limit：因为数据世系的个数与其规模呈幂率分布，所以只有少数数据世系含有大量的消息。此处，令数据世系的消息上限为固定值(Bundle Limit)，同时仍采用上述的 Partial Index。由于采用了 Bundle Limit 方法，使得内存中的数据世系个数增多了，能更好地应对用户的查询。

令 E_0 为采用 Full Index 方法得到的数据世系的边，令 E_1 为采用 Partial Index 或者 Bundle Limit 得到的边，那么 Partial Index 或者 Bundle Limit 方法的准确率为

$$accuracy_1=\frac{|E_1\cap E_0|}{E_1} \quad (10)$$

3.2 数据世系的发现性能

图 2 为 3 种算法得到的数据世系的个数随着消息增加的变化图。Full Index 方法得到的数据世系个数随着消息数量的增长呈线性增长，当大量的微博消息流进入系统时，该方法显然不能在内存中保存所有的数据世系。Partial Index 和 Bundle Limit 两种方法的数据世系个数先增加，当文本流趋于稳定时，数据世系的个数变化很小，因此可用于微博流数据的实时处理。图 3 为 Partial Index 和 Bundle Limit 两种方法的准确性对比图。当新消息到达时，将该消息与内存上的数据世系进行匹配，同时将其加入到最合适的数据世系中。当新消息不断增多时，这两种方法匹配得到的数据世系个数先增加，然后趋于恒定值。对于准确性，Partial Index 要高于 Bundle Limit。

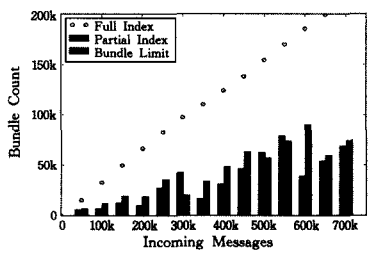


图 2 数据世系的数量随着消息数量的变化情况

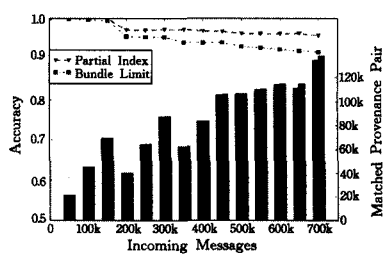
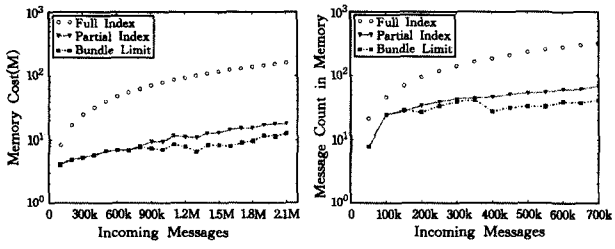


图 3 算法的准确性对比

3.3 内存开销与运行效率

Full Index 方法虽然用于离线计算，但是当数据流的规模比较小时仍可以通过内存方法来实现。图 4 为 3 种算法随着消息个数不断增多时的内存开销对比图。图 4(a)为内存的使用量，图 4(b)为内存中保存的消息个数。从图 4 中可以看出，Full Index 方法在消息个数增加时，内存的使用量和内存中的消息个数都快速增加，因此不适用于内存算法，不能用于微博数据流的实时处理。Partial Index 和 Bundle Limit 方法的内存开销虽然也随着消息个数的增长而增长，但是增长缓慢，并且当消息个数到达一定数量后，可以通过将过时的数据世系存放到磁盘上来保持内存使用的平衡。



(a)内存使用　　(b)消息个数

图 4　算法的内存开销对比

最后，实验对比了 3 种算法的运行效率，结果如图 5 所示。随着消息数量的增长，3 种方法的运行时间都近似线性地增长。由于 Partial Index 方法具有较低的内存开销，因此其数据世系的发现性能要优于其它两种方法。

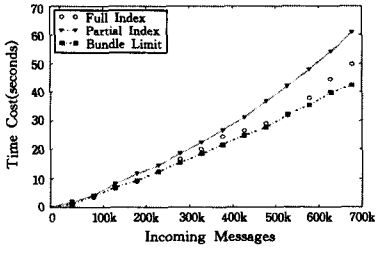


图 5　算法的运行效率对比

结束语　微博信息检索是微博研究的重要内容之一。微博平台中的消息以流的形式源源不断地流入系统，这些消息间按照事件的起源、发展和变化具有一定的关联性。本文基于数据库中的数据世系思想，根据消息中包含的内容以及各种指示符将同一事件包含的消息划分为一个数据世系。于是文本数据流变成了数据世系集合的维护。随着时间的推移，如果某事件不再被关注，那么从内存中移除该数据世系，以便为新的数据世系提供内存空间。当用户检索时，微博的即时性(检索的内容往往是近期的)使得响应的结果往往包含于内存中。实验表明，基于数据世系的微博消息管理方法使用的内存少，运行效率高，可用于微博消息流的实时处理及查询响应工作。

参 考 文 献

[1] Miller G. Social scientists wade into the tweet stream [J]. Science, 2011, 333(6051): 1814-1815

[2] Java A, Song X, Finin T, et al. Why we twitter: understanding microblogging usage and communities[C]// Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, 2007: 56-65

[3] Oh C, Sheng O. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement[C]// ICIS. 2011

[4] Sprenger T O, Tumasjan A, Sandner P G, et al. Tweets and trades: The information content of stock microblogs[J]. European Financial Management, 2014, 20(5): 926-957

[5] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data[C]// Proceedings of the Workshop on Languages in Social Media. 2011: 30-38

[6] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848
Zhao Yan-yan, Qin Bing, Liu Ting. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848

[7] Gaonkar S, Li J, Choudhury R R, et al. Micro-blog: sharing and querying content through mobile phones and social participation [C]// Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services. ACM, 2008: 174-186

[8] 刘志明，刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程，2011，29(6)：8-16
Liu Zhi-ming, Liu Lu. Recognition and Analysis of Opinion Leaders in Microblog Public Opinions[J]. Systems Engineering, 2011, 29(6): 8-16

[9] Simmhan Y L, Plale B, Gannon D. A survey of data provenance in e-science[J]. ACM Sigmod Record, 2005, 34(3): 31-36

[10] Moreau L. The foundations for provenance on the Web [J]. Foundations and Trends in Web Science, 2010, 2(2/3): 99-241

[11] Budak D, Abbadi A E. Information diffusion in social networks: Observing and influencing societal interests[J]. PVLDB, 4(12): 1-5

[12] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle[C]// Proc. of KDD. 2009: 497-506

[13] Wu S, Hofman J, Mason W, et al. Who says what to whom on twitter[C]// Proc. of WWW. 2011: 705-714

[14] Teevan J, Ramage D, Morris M R. # twittersearch: a comparison of microblog search and Web search[C] // Proc. of WSDM. 2011: 35-44

[15] Chen C, Li F, Ooi B C, et al. Ti: An efficient indexing mechanism for real-time search on tweets[C] // Proc. of SIGMOD. 2011: 649-660

[16] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models[C]// Proc. of ICWSM. 2010

[17] Zhao X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]// Proc. of ACL. 2011: 379-388

[18] 李波，石慧霞，王毅. 一种基于同义词发现的文本扩充算法[J]. 重庆理工大学学报(自然科学版)，2014，28(2)：76-81
Li Bo, Shi Hui-xia, Wang Yi. A Text Extension Algorithm Based on Synonymy Discovery[J]. Journal of Chongqing University of Technology(Natural Science), 2014, 28(2): 76-81