面向主题的社交网络采集技术①

郑楷坚, 沙 瀛

(中国科学院 信息工程研究所, 北京 100093)

摘 要: 社交网络数据采集是开展社交网络分析的基础. 针对当前面向主题的社交网络数据采集技术采集数据少、召回率低的问题, 本文提出基于内置搜索引擎和基于通用搜索引擎相结合的主题消息采集方法,并将LDA(Latent Dirichlet Allocation,隐含狄利克雷分布)模型应用于主题关键词的迭代扩展, 并提出了一种基于用户生存值的高效扩展策略. 实验结果表明本文提出的方法可以使面向主题的社交网络数据采集系统在保证一定准确率的情况下进一步获取主题相关数据.

关键词: 社交网络; 主题采集; 内置搜索; 通用搜索; 主题模型; LDA

Topic Focused Crawling Technique on Social Network

ZHENG Kai-Jian, SHA Ying

(Institution of Information Engineering, Chinese Acadamic of Sciences, Beijing 100093, China)

Abstract: Social network data is the basis of social network analysis that is why it's important to collect such data. To solve the problem of less collected data and low recall rate in current focused crawlers on social network, this paper proposes a method combining the based built-in search engine and general search engines to crawl topic messages, as well as applys the LDA model to extract the topic keywords from collected data iteratively and adds new topic keywords to the seed. Besides, an efficient expansion strategy based on users survival value is discussed. Our experiment shows that the methods proposed can improve the recall rate with a high precision.

Key words: social network; focused crawler; built-in search engine; general search engine; topic model; LDA

1 引言

互联网技术的进步和智能手机的普及促进了社交网络的蓬勃发展. 根据统计, Facebook 的月活跃用户在 2015 年第三季度便突破 15 亿, Twitter 的月活跃用户数也达到 3 亿. 在移动端方面, Mary Meeker 发布的《2015 年互联网趋势报告》[1]显示, 用户量最多的前十大移动端应用中社交类应用占据 6 席.

社交网络用户规模巨大, 内容丰富, 影响范围广, 具有十分重要的研究意义. 然而, 由于社交网络数据为服务商私有, 通过采集手段获取社交网络数据便成为重要途径. 当前, 社交网络采集技术主要有基于官方 API[2]、基于浏览器[3,4]和基于模拟 AJAX[5]三种方式. 其中, 基于官方 API 的方式, 其采集的内容准确、较全面, 带宽占用小, 但是受限严格; 基于浏览器的方式实现简单, 适用性强, 但系统资源消耗较大, 采集速度慢; 基于 AJAX 模拟的方式, 采集速度快, 内容全面, 但是开发难度大, 维护成本高.

日益复杂庞大的社交网络及其海量内容数据对社交网络数据的采集带来了巨大的挑战. 由于网络连通性难以保证等原因, 全网数据的采集难以实现, 而且社交网络的动态性决定了我们通常只能获取每一时刻或者多个时刻的网络快照. 当前的社交网络研究工作主要集中在特定主题的数据集的分析. 因此, 面向主题的社交网络采集技术研究具有重要性和必要性.

当前, 许多面向主题的采集策略和方法被提出并应用到具体的任务中. 但是, 这些方法普遍存在着采

① 基金项目:国家科技支撑计划(2012BAH46B03)

收稿时间:2016-01-28;收到修改稿时间:2016-04-29 [doi:10.15888/j.cnki.csa.005383]

集回来的数据少，召回率较低的问题. 大量的社交网络事件监测系统在实现主题消息的采集时严重依赖于社交网络内置的搜索引擎，数据的质量和数量无法保证. 另一方面，主题的表示固定不变，没有根据已采集的主题数据进行迭代更新，容易导致主题新数据的丢失. 比如对于主题“环境污染”，在“柴静发布雾霾调查”事件发生后，社交网络上会有很多相关的消息，如果不对采集回来的部分新数据进行分析学习，这个子主题的很多消息都将采集不到. 已有的针对主题用户的采集系统，通常只从主题用户扩展，并且缺少对内置搜索引擎和元搜索的利用，很难全面地获取主题用户.

针对上述的问题，本文从三个方面进行了改进. 在主题消息的采集方面，将 基于内置搜索和基于元搜索的方式相结合，提升了采集的覆盖率，可以获取更多主题消息，同时这种方法并不只针对某个社交网络，对大多社交网络都适用. 主题关键词的扩展方面，每次采集到一定主题数据后，利用 LDA 模型进行主题关键词的抽取，将潜在的或新的主题关键词扩展到主题表示里，加入采集[6]. LDA 模型是一种主题模型，可以从文档集中学习到”文档-主题”分布以及”主题-单词”分布，在学习到”主题-单词”分布后，便可实现主题关键词的抽取. 主题的迭代扩展，一方面可以实现主题新数据的获取，另一方面，可以降低原始主题表示不全面所带来的影响. 针对主题用户的采集，提出新的扩展策略. 通常的采集系统在计算完相似度，发现用户不是主题用户后，便停止从用户处扩展，很容易出现主题用户的遗漏. 本文给每个用户设置一个生存值属性 ttl(time to live)，只有用户的 ttl 值为 0 时，才不进行扩展. 当用户为主题用户时，由其扩展的所有用户 ttl 值为为前用户 ttl 值加 1，如果不是主题用户，则扩展的用户 ttl 值为当前用户的 ttl 值减 1. ttl 值同时是用户采集的优先级，ttl 值高的用户，优先进行采集. 同时，利用内置搜索和元搜索实现初始主题用户的选取，减少人工的干预. 相比于无目的的滚雪球策略，这种方式采集效率更高，与采用严格扩展策略的方法相比，本文的方法又有更大的可能性发现遗漏的主题用户.

本文的结构安排如下：第 2 节对相关的研究工作进行了总结，主要是当前面向主题的社交网络采集的分析. 第 3 节具体介绍了本文提出的面向主题的社交网络数据采集技术，详细阐述了主题消息采集中基于内置搜索引擎与基于通用搜索引擎相结合的方式、利用 LDA 模型实现主题关键词扩展的方法以及主题用户采集中使用的基于用户生存值 ttl 的一种高效扩展策略. 第 4 节介绍了本文最终实现的针对 Twitter 的面向主题的社交网络数据采集系统. 第 5 节是实验结果及分析. 最后第 6 节是对本文的总结.

2 相关研究工作

面向主题的社交网络采集，是指从社交网络上采集与特定主题相关的社交网络数据. 采集系统首先读取种子，即系统的输入，一般为用户或者关键词. 在采集完某个用户或者关键词的数据后，结合主题表示和相似度计算方法，比如基于词匹配的方法或者训练好的分类器对数据的主题相关性进行判断. 主题相关的数据保存到数据库中. 对于用户的采集，最后还会根据相似度计算结果和扩展策略决定是否将用户的社交网络关系加入待采集行列.

与一般的社交网络采集相比，面向主题的社交网络采集在定制种子时，需要更多的主题相关知识，才能保证主题定义的准确性和全面性. 与主题紧密相关的还有相似度计算，计算方法的选择和计算结果的精度很大程度上决定了面向主题的采集系统的效率和准确性.

主题的定义和描述上，基于关键词的主题表示是最简单方法，在该方法上通过在关键词之间添加”and”和”or”的关系可以得到基于规则的主题表示，即主题由多条规则表示，每条规则由多个关键词之间通过”and”或者”or”连接. 仲兆满等人利用通用搜索引擎和内置搜索引擎进行采集，通过将主题规则分解为原子规则，并使用句群进行相关性过滤，提升了面向主题采集的准确率[7]. 由于没有考虑语义信息，在一词多义或者一义多词的情况下，利用基于关键词或者基于规则的表示方法进行相似度计算，容易引进噪音信息或者发生遗漏. Dong H 等人通过对主题构建领域主体，补充了语义信息，提升了相似度计算的准确性[8].

Matko 利用用户的地理位置信息和推文的语言信息进行相关性的判断，并从用户的粉丝关系进行扩展，实现了 twitter 上葡萄牙用户的采集[9]. Chi-In 利用浏览器内核 WebKit，同样通过用户的地理位置信息进行判断，实现 Facebook 上澳门用户的采集[10]. Andrei 基于知识库训练出文档分类器用于相似性判定，同时利用

非齐次柏松过程对用户每天的博客数目进行建模，齐次柏松过程对博文发表时间进行建模，从而实现对目标主题数据的高效实时采集[11].

3 面向主题的社交网络采集技术改进

3.1 主题消息采集

社交网络主题消息的采集通常利用搜索引擎实现. 大部分社交网络都会有内置的搜索引擎，便于用户搜索站内的信息. 因此，已有的面向主题的社交网络采集系统在进行主题消息采集时都将其作为第一选择. 但是，不管以何种技术具体实现，这种方法都存在着明显的不足. 首先，返回的结果受到后台的严格控制，数量和质量无法保证. 其次，不同社交网络对搜索功能的支持程度存在较明显的差距，比如 Facebook 只支持对用户和公共主页的检索，Twitter 则支持对用户、推文、图片和视频等的搜索，因此，单独使用内置搜索引擎的方法，适用性不高.

另外一个可以利用的是通用搜索引擎，即借助 Google、Bing 和 Yandex 等支持站内搜索的搜索引擎，将关键词发给各搜索引擎，对返回的结果进行去重整理. 与内置搜索引擎相比，通用搜索引擎返回的数据结果多，但也包含了许多不相关的其它类型的数据. 以搜索 Facebook 上的”旅行”为例，返回的结果有用户、消息、活动、小组等各种类型. 另外，在需要对主题数据实时获取的场景下，通用搜索引擎并不适用.

针对两种搜索引擎存在的问题，本文将两者结合起来. 在读取种子输入之后，同时利用两者进行搜索，通用搜索引擎的部分结果会反馈于内置搜索引擎. 过程如图 1 所示.

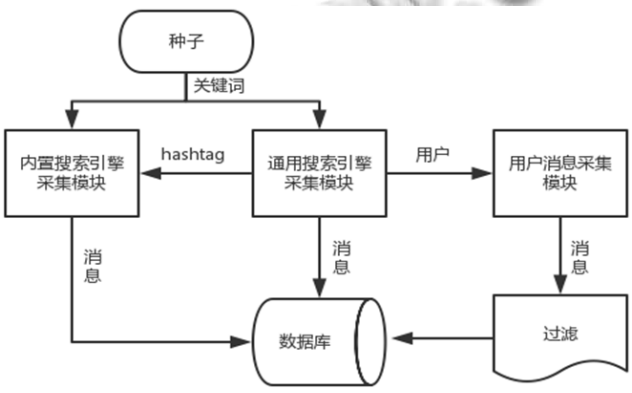

图 1 关键词采集示意图

对于内置搜索引擎，如果支持关键词消息检索，则直接利用关键词进行检索，并将采集到的消息数据保存到数据库，如果内置搜索引擎只支持关键词用户检索，则对检索出来的用户，再进一步调用采集系统的用户消息采集模块进行采集，并根据关键词进行主题消息的过滤，将主题数据保存到数据库.

对于通用搜索引擎，由于大部分的社交网络都有消息、用户和 hashtag 这三个类型，因此，在分析和抽取采集结果时只对这三种类型进行处理:

① 如果是消息类型，则将数据直接保存到数据库;

② 如果是用户类型，当采集系统支持用户消息采集时，从数据中抽取出用户 ID，并调用用户消息采集模块进行采集，过滤出命中的消息保存到数据库;

③ 如果是 hashtag，则将 hashtag 从数据中抽取出来，并将 hashtag 作为新的关键词放入到内置搜索引擎进行采集.

3.2 主题关键词扩展

在主题由一组关键词表示的情况下，关键词选取的准确性和全面性对于采集的效果影响很大. 当主题范围比较大或者专业性比较强时，关键词的选取将很困难. 另一方面，希望采集系统能够有一定的学习能力，可以在采集的过程中发现新的主题关键词，并作为新种子加入采集，提升采集的覆盖率. 基于上述两个方面的考虑，每次采集完一定的主题数据后，本文利用 LDA 模型进行主题关键词的提取，实现对主题关键词进行扩展.

LDA 模型是经典的主题模型. 记“文档-主题”分布为 θ，“主题-单词”分布为 Φ，LDA 模型定义的文档生成过程如下:

① 根据 θ，选择一个主题，记为 t;

② 根据 Φ 为主题 t 选择一个单词，作为文档的单词;

③ 重复 1、2 过程直到文档生成.

在给定文档集的情况下，通过反复实验等方法得到主题数目 T，再利用变分-EM 算法或者 Gibbs 抽样法便可以实现 θ 和 Φ 的推断. 通过这两个参数的学习，我们可以获得文档中的主题以及每个文档所涵盖的主题比例. 由于 LDA 可以学习到每个主题的词项分布，可以利用 LDA 主题模型进行主题关键词抽取.

LDA 的学习需要先预估主题的数目，如果直接将主题数设置为 1 进行学习，并没有意义. 在实际当中，

一个主题通常会包含多个子主题，子主题之间有一定的相似性，但也存在着较大的差异．以主题“环境污染”为例，“柴静雾霾调查”和“天津港爆炸污染”可以视为两个子主题，它们都涉及到空气污染，但是其实是两个相对独立的事件．基于上述的考虑，本文将对一个主题的学习转化为对多个子主题的学习过程．得到子主题的词项分布之后，再综合计算每个词项在主题的权重值．

在词项权重计算完毕后，如果直接将权重最大的 K 个词作为关键词进行采集，采集的误差将很大．依然以“环境污染”为例，“北京”这个词必须会有较大的权重．如果直接采集“北京”，将把很多政治主题的推文采集回来．因此，在扩充的时候，先将主题关键词进行两两配对，词组对的权值为原来两个关键词各自权值的和，最后选取权值最大的 k 对词组．通过这种方式，加入采集的将是“北京 雾霾”、“空气 污染”这样的词组．主题下的子主题数目随着时间的变化，通常会增加．因此，在每轮进行 LDA 学习时，子主题数目应该比上轮的数目要多．在本文的方法中，每轮的子主题数目加设为上一轮子主题数目加 1．LDA 提取关键词并进行扩充的整个过程如下:

① 设置子主题数目 num;

② 利用 LDA 学习各个子主题的词项分布;

③ 对每个词项 t_i，根据其在所有子主题的分布权重，得到其主题权重 w_i;

④ 对每两个词项 t_i, t_j，计算 $w_i + w_j$，将该值最大的前 k 对作为新的主题关键词加入种子．

3.3 主题用户采集

主题用户的采集需要高效的扩展策略．已有的扩展策略在采集效率和采集的覆盖率两方面很难做到兼顾．针对这个问题，本文提出了一种扩展策略用于采集更多的主题用户．

首先，给用户任务队列里面的每个用户一个生存值 ttl．初始种子用户的 ttl 值设为 t，t 可以根据采集的具体需要设为任意大于 0 的整数．对于当前正在采集的用户，首先采集其消息数据和个人信息数据，利用这些数据计算其主题相似度．如果最终主题相关，则进一步采集用户的社交网络关系，将其好友粉丝作为待采集用户加入任务队列，新用户的 ttl 值为当前用户的 ttl 值加 1．如果主题不相关，在当前用户的 ttl 值已经降为 0 时，则不再扩展，否则，依然扩展用户的好友粉丝到任务队列中，此时，这些用户的 ttl 值为当前用户的 ttl 值减 1．任务队列是一个优先级队列，ttl 值高的用户优先采集．使用 ttl 值的目的是为了让扩展不过早停止，减少遗漏的可能．使用优先队列是为了让主题相关可能性更大的用户提前采集，提升效率．具体的扩展流程如图 2 所示．

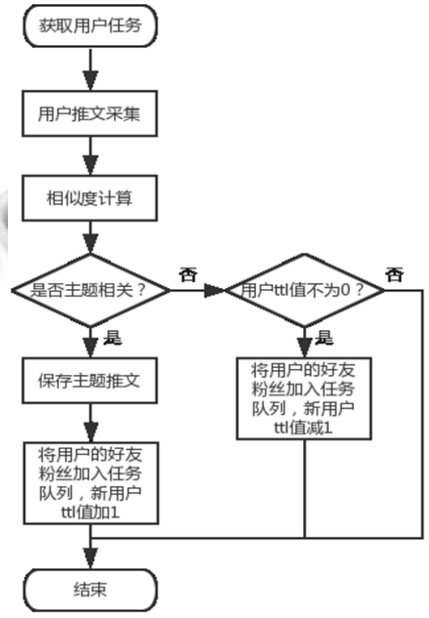

图 2　用户任务扩展流程

初始种子用户的选取，可以利用通用搜索引擎和内置搜索引擎，从而减少人工干预．

4　面向主题的社交网络数据采集系统

针对 Twitter 我们实现了面向主题的社交网络数据采集系统．系统架构如图 3 所示．

系统主要由四个部分组成，分别为采集服务器、消息队列、采集节点和内容计算节点．采集服务器的主要功能是负责采集任务的生成、系统状态和采集进度的管理和监控．消息队列用于采集服务器和采集节点之间的通讯．采集服务器和采集节点之间不直接交换数据，而是以消息的形式放到对应的队列当中，消息接收方在取得消息后，根据一定的规则去解析消息，并执行相应的动作．采集节点负责具体的采集，它的主要功能模块包括内置搜索引擎采集模块、通用搜索引擎采集模块、用户消息采集模块、用户好友粉丝采集模块以及用户个人信息采集模块．内容计算节点完成采集数据的主题相似度计算和主题关键词的扩展．

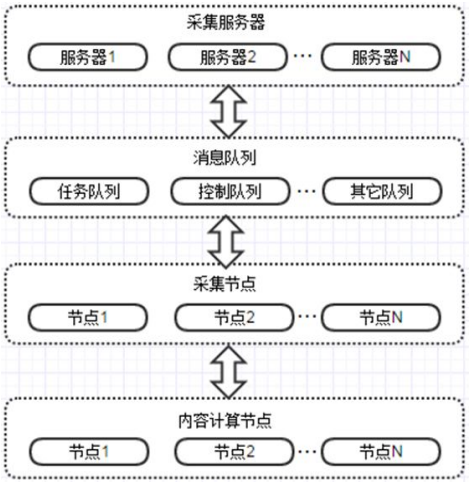

图 3 采集系统架构图

5 实验结果及分析

本文的实验主要包含两部分: 1) LDA 模型在关键词扩展中的效果; 2) 本文提出的三种方法在面向主题的采集系统中的应用效果.

为了验证 LDA 模型在关键词扩展中的效果, 本文选取基于 TF(Term Frequency, 词项频率) 和 DF(Document Frequency, 文档频率)的两种关键词扩展方法进行对比, 同时定义了准确率 P_m 作为评价指标, 具体如下:

$P_m = n_{rm}/n_{cm}$, 其中, n_{rm} 表示实际的主题消息数, n_{cm} 表示实验中采集回来的主题消息数.

本次实验的主题设置为“环境污染”, 即种子输入为关键词“环境污染”. 此次实验总共进行 4 次主题关键词扩展, 每次扩展添加 5 个新的关键词组到种子中. LDA 的初始子主题数设为 10. 其中, 第一轮主题关键词扩展后新添加的 5 个关键词组分别如下:

LDA: “环境 污染”、”污染 中国”、” 污染 癌症”、“污染 环保”、”污染 空气”

TF: “环境 污染”、”污染 中国”、”污染 严重”、” 污染 经济”、“污染 北京”

DF: “环境 污染”、“环境 中国”、“环境 严重”、“环境 问题”、“环境 经济”

从每次迭代后采集回来的消息中随机取 200 条进行人工的主题相关性标注, 并计算准确率 P_m, 其结果如图 4 所示.

从实验的结果可以看出, 在 4 次扩展中, 基于 LDA 的关键词扩展方法始终保持着较高的准确率, 比较稳定.

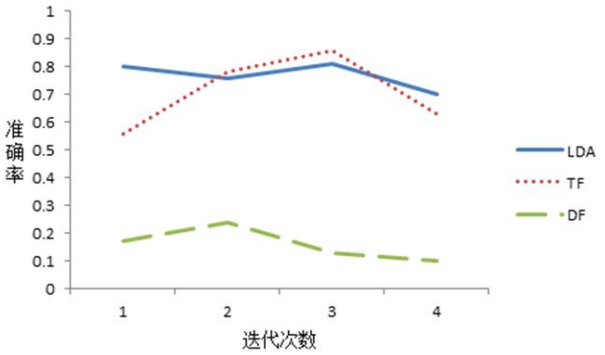

图 4 三种关键词扩展方法的准确率

接着, 为测试本文所提出的三个方法的应用效果, 我们实现了 4 个面向主题的采集系统进行实验对比, 以检验这三个方法是否能改进面向主题的社交网络采集的技术.

第一个系统, 基于滚雪球扩展策略, 主题消息采集只集成了内置搜索引擎采集模块, 记为 S1.

第二个系统, 基于本文提出的扩展策略, 同样只集成了内置搜索引擎采集模块, 记为 S2.

第三个系统, 基于本文提出的扩展策略, 主题消息采集使用我们提出的内置搜索引擎和通用搜索引擎相结合的方式, 记为 S3.

第四个系统, 在第三个系统的基础上, 加上本文提出的关键词扩充技术, 记为 S4.

对这 4 个系统, 在给定同样输入的情况下, 分别用以下 4 个指标进行对比:

① 主题消息规模;

② 主题用户规模;

③ 主题消息的准确率:

$P_m = n_{rm}/n_{cm}$, 其中, n_{rm} 表示实际的主题消息数, n_{cm} 表示实验中采集回来的主题消息数.

④ 主题用户的准确率:

$P_u = n_{ru}/n_{cu}$, 其中, n_{ru} 表示实际的主题用户数, n_{cu} 表示实验中采集回来的主题用户数.

在实验中, 这 4 个系统所计算出来的各指标结果分别如图 5-8 所示.

① 图 5 是主题消息规模的对比结果, 其给出了 S1-S4 这 4 个系统在前 4 个小时内的主题消息规模. 可以观察到 S4 表现出明显的优势, S2 略高于 S1, 而 S3 又比 S2 进一步提升了规模.

② 图 6 是主题用户规模的对比结果, 其给出了这 4 个系统在前 4 个小时内的主题用户规模. 可以看到,

这 4 个系统的变化趋势跟主题消息基本一致.

③ 图 7, 图 8 分别是主题消息准确率 P_m 和主题用户准确率 P_u 的对比结果. 计算 P_m 和 P_u 时, 对每个系统, 分别从每个时间段内随机抽取出 200 条消息进行主题消息标识, 抽取出 100 个主题用户进行主题用户标识, 若用户发布过至少一条主题相关的消息, 便标识其为主题用户. 由图 8, 图 9 可以观察到, 这 4 个系统的准确率基本维持在同一个等级, 但 S1、S2 的准确率略高于 S3、S4. 这主要是因为通用搜索引擎采集到的结果越往后, 其相关性越较低, 关键词扩展部分越容易引入非相关的关键词组.

由上述的实验结果可以看出本文所提出的面向主题的社交网络采集的三个方法能够在保持准确率的情况下, 提高采集的覆盖率和规模. 其中, 关键词扩展由于补充了输入而效果最为明显. 内置搜索引擎和通用搜索引擎相结合的方式因为补充了采集通道, 采集数量有所提高, 采集扩展策略对面向主题的采集也有一定的帮助.

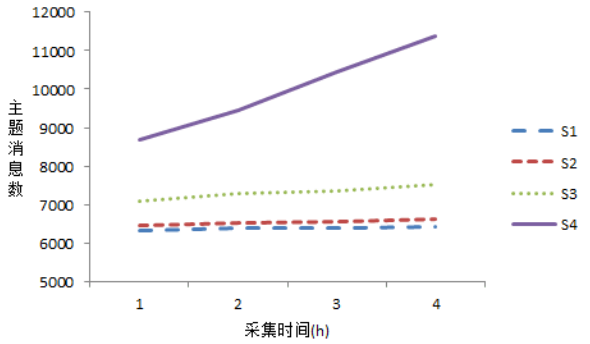

图 5 S1-S4 的主题消息规模

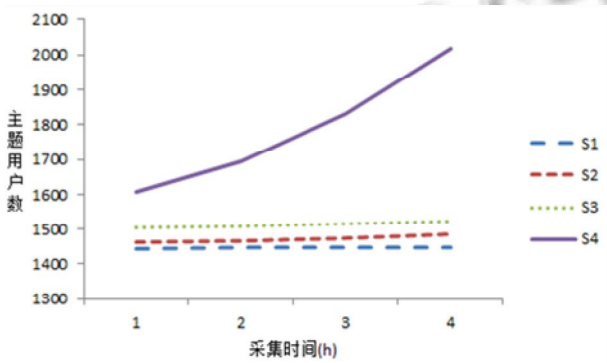

图 6 S1-S4 的主题用户规模

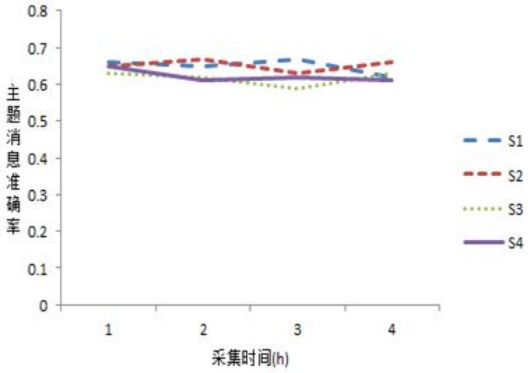

图 7 S1-S4 的主题消息准确率

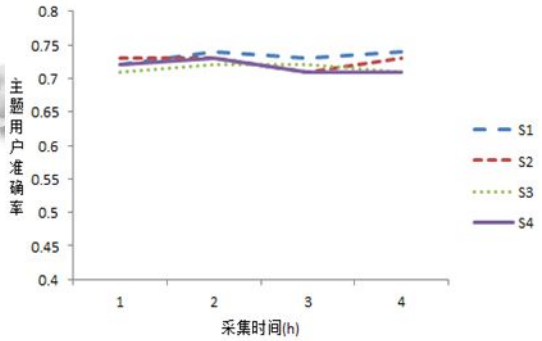

图 8 S1-S4 的主题用户准确率

6 结论

针对当前面向主题的社交网络采集存在着采集覆盖率低, 采集效率差的问题, 提出了将内置搜索引擎和通用搜索引擎相结合的主题消息采集方法, 同时利用已采集的主题数据进行主题关键词的迭代抽取, 扩充种子, 并改进了采集过程中的扩展策略. 实验结果表明, 本文设计的面向主题的社交网络采集系统能很好地完成 Twitter 上主题推文和主题用户采集的任务, 在保持一定采集精度的情况下, 其在主题采集速度和采集覆盖率上比普通主题采集系统都有较大提升. 下一步将引进语义信息, 对采集数据与主题相似度的计算方法进行改进, 同时提升采集系统的通用性.

参考文献

1 Meeker M. Internet trends 2015 – code conference. Glokalde, 2015, 1.3.

2 房伟伟,李静远,刘悦,等.Twitter 数据采集方案研究.山东大学学报(理学版),2012,47(5):73–77.

3 陈飞.基于 WebKit 浏览器引擎的动态页面数据采集方案. 中国科技论文在线. 2010. http://www.paper.edu.cn/releasepaper/content/201012-730.

4 陈学敏,沙瀛.基于浏览器测试组件的社交网络数据获取技术研究.信息网络安全,2015,(5):56–61.

5 单既喜.社交网络数据获取及其分析系统[学位论文].北京:中国科学院大学, 2013.

6 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993−1022.

7 仲兆满,李存华,刘宗田,等.一种基于搜索策略的多主题信息采集方法.电子学报,2013,42(12):2352–2358.

8 Dong H, Hussain FK, Chang E. A transport service ontology-based focused crawler. Fourth International Conference on Semantics, Knowledge and Grid, 2008. SKG'08. IEEE. 2008. 49–56.

9 Boanjak M, Oliveira E, Martins J, et al. Twitterecho: a distributed focused crawler to support open research with twitter data. Proc. of the 21st International Conference Companion on World Wide Web. ACM, 2012: 1233–1240.

10 WONG CI, et al. Design of a Crawler for Online Social Networks Analysis. Wseas Trans. on Communications, 2014, 13: 263−274.

11 Yakushev AV, Boukhanovsky AV, Sloot PMA. Topic crawler for social networks monitoring. Knowledge Engineering and the Semantic Web. Springer Berlin Heidelberg, 2013: 214–227.