文章编号:1671-9352(2012)03-0000-00

基于语义分析的微博搜索

刘晓华[1,2],韦福如[2],段亚娟[3],周明[2]
(1.哈尔滨工业大学计算机科学与技术学院,黑龙江 哈尔滨 150001;
2.微软亚洲研究院,北京 100080;3.中国科技大学计算机科学与技术学院,安徽 合肥 230026)

摘要:提出构建基于语义分析的微博搜索以帮助用户从海量的、书写通常不规范的微博中有效地获取信息。定义了微博语义搜索要解决的问题,讨论了微博语义搜索所面临的挑战及对策,介绍了一种参考实现框架及相关的语义分析技术,特别是面向微博的语义角色标注技术。

关键词:微博;搜索引擎;语义搜索;语义角色标注
中图分类号:TP391.3 **文献标志码**:A

Semantic search of micro-blogs

LIU Xiao-hua[1,2], WEI Fu-ru[2], DUAN Ya-juan[3], ZHOU Ming[2]
(1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China;
2. Microsoft Research Asia, Beijing 100080, China; 3. School of Computer Science and Technology,
University of Science and Technology of China, Hefei 230026, Anhui, China)

Abstract: A search engine based on semantic analysis for micro-blogs (semantic search of micro-blogs) is proposed, Aiming to the efficent access of information from the huge number of micro-blogs (such as tweets) which are short and often informally written. Unlike current micro-blogs search engines, it conducts serials of natural language processing and text mining for micro-blogs to get interesting points such as named entities, events and opinions, which are further indexed, and thus enables two brand new scenarios, i.e., categoriezed browsing and advanced search. The challenges and their possible solutions, a reference implementation framework, and related core semantic computing technologies, e.g., semantic role labeling, are presented as well.

Key words: Micro-blog; search engine; semantic search; semantic role labeling

0 背景介绍

最近3年,微博服务迅速流行[1-2]。例如,国外最大的微博服务提供商推特(Twitter:http://www.twitter.com)目前每天产生超过5 500万的微博(字数不超过140字符的短文本),用户访问数超过1.8亿,其注册用户则超过了1亿。国内微博服务也开始兴起。目前主流的互联网门户都提供了微博服务。其中比较典型的代表是新浪微博。截止2011年4月它已经拥有超过1.4亿的微博注册用户[3]。微博服务之所以流行,在于它满足了草根快速交流分享信息的需要。在微博平台上,任何用户可以就任何话题发布任何消息,此外,微博用户还可以追随其他用户,或者向追随者推荐微博,使得微博平台成为一个巨大的社会化网络。巨大的信息量和庞大的用户群体,使得微博已经成为一个最重要的实时信息源,一种影响力日益增强的新的社会媒体,一些重要的热点事件,都是由微博首先报道的,例如,2009年迈克·杰克逊逝世的消息、2010年智利大地震以及2011年的本·拉登被击毙等。

收稿日期:2011-11-30;网络出版时间:2012-03-20 10:59
网络出版地址:http://www.cnki.net/kcms/detail/37.1389.N.20120320.1059.017.html
作者简介:刘晓华(1976-),男,研究员,博士,研究方向为自然语言处理. Email:Lxh5147@126.com

为了帮助用户从海量的,动态更新的微博中找到感兴趣的内容,推特提供了微博搜索服务。和传统的搜索引擎类似,推移用户输入关键字进行搜索,就可以得到了按时间排序的微博列表。尽管其搜索功能相对简单,但其访问量巨大。据报道,自2009年4月份以来,推特搜索的次数增长了33%,现在推特每天处理的搜索请求次数超过了8亿次,超过了微软必应和雅虎的检索请求之和[4]。

传统搜索引擎厂商已经认识到微博对搜索的重要性,并推出了各自的微博搜索。例如微软必应(http://www.bing.com)购买了推特微博数据的授权,提供了相应的搜索服务(http://www.bing.com/social),一批创业公司也纷纷推出了基于微博的搜索,例如twazzu(http://www.twazzup.com)。

1 现有微博搜索的不足

首先,目前的微博搜索只提供了基于关键字的搜索接口,无法满足用户快速从微博中获取信息的需要。这里面有两个深层原因:一方面,因为微博的量大,简单的关键字匹配会返回很多结果,甚至会包括多条几乎相同的微博;另一方面,因为微博短,单条信息不全,而且其写法随意,所以用户在阅读以列表方式显示的微博搜索结果时感觉比较吃力。

其次,因为目前的微博搜索对微博内容本身的理解仅停留在关键词阶段,所以用户需要花费大量精力对搜索结果进行整理提炼,才可能得到所要的信息。例如,要了解最近微博中热门的事件,热门的人物,热门的观点等信息,仅靠目前的微博搜索是很困难的。

最后,现有的微博搜索缺乏对商业智能的支持。微博服务平台的一个重要价值就在于它反映了草根的声音,而草根的声音蕴含着相当大的商业价值。例如,公司在发布新产品后,会非常在意大众对其的评价,以便采取针对性的市场措施或产品改进。而现有的微博搜索,不能支持用户对某个关注对象,例如新产品,正负面评价的概况。

总之,当前的微博搜索基本上是传统网页搜索的简单克隆,缺乏对微博内容的挖掘整理,因而只能对用户仅提供非常有限的帮助。

2 主要任务

我们提出构建基于语义分析的微博搜索(简称微博语义搜索)解决现在微博搜索面临的问题。区别于现有的微博搜索,微博语义搜索强调对微博内容做深入挖掘和系统整理以获得结构化的信息点(例如事件、情感),并基于这些高度结构化信息提供导航和搜索服务。具体而言,语义微博搜索的主要任务是解决如下4个问题:

第一,如何从微博中提取用户感兴趣的信息点?我们定义3类信息点:实体,包括人物、机构、地点、时间;事件,也就是谁在什么时间什么地点以什么方式对谁做了什么事;情感,也就是谁对谁持有正面或负面的评价。

第二,如何让用户方便的访问这些信息点,而不是仅通过简单的搜索?我们定义分类浏览和搜索两种方式。分类采用和主流新闻搜索一致的分类体系,也就是类似科技、政治、娱乐、生活、教育、商业、健康等这些类别;除按关键字搜索外,还支持按事件或情感搜索。例如,能搜索“和比尔.盖茨相关的事件”(“events involving Bill Gates”)或“关于奥巴马的负面意见”(“negative opinions about Obama”)。

第三,如何对搜索结果进行有序化整理,消除冗余的或者垃圾微博?我们定义搜索结果聚类操作,把内容相同或相似的微博组织到一起,并只为每个聚类只显示一个有代表性的微博。

第四,如果支持商业智能?按情感搜索允许我们在某种程度上支持商业智能。此外,对每个输入的查询,如果该查询表示一个实体,情感概览操作将执行,也就是显示正面、负面意见的比例,并且为两类意见给出代表性的表示情感的词作为摘要。

3 挑战及对策

首先,微博数量巨大,时效性强。这要求微博语义搜索处理速度快,支持实时索引。在这一点上,和传统搜索引擎迥然不同。例如,目前最流行的谷歌网页搜索引擎的更新周期平均约2周,而微博语义搜索的更新速度要达到秒级。也就是说微博发布后,从进入语义搜索平台进行信息抽取,到索引就绪,总共花的时间应在10 s以内。

对于此,主要的应对措施包括:仅关注最新的微博,例如最近一周的微博,这样可大大缩小索引的数据量;设计灵活的索引结构,并对批量写操作进行特殊优化;采用简单高效的方法实现信息抽取;考虑到微博的海量性,重点保证抽取信息点的准确率,适当降低召回率。

其次,微博中相当一部分是对一般用户毫无价值的垃圾信息。据报道,推特上40%的微博对一般

用户而言是无意义的[5]。如果不对这些垃圾做清理,一方面会浪费大量的计算和存储资源,另一方面也会影响信息点抽取的质量。

针对于此的对策包括:开发专门的过滤器,在微博进入到语义搜索系统前,用这些过滤器过滤掉疑似垃圾微博;通过对检索结果进行聚类,并只显示前 N(例如 10)个有代表性的微博,也可以在一定程度上帮助过滤垃圾微博。

最后,因为单条微博表达的意思一般不完整,并且通常微博书写都较随意,从而使得从微博中抽取信息点的难度比从正规文本,例如新闻,抽取的难度大。与此相关的另一个后果是:原来工作良好的处理工具,例如词性标记工具,命名实体抽取工具等,直接应用到微博上时的效果差。因为这些工具基本上都是在和微博有着巨大差异的正式风格的文本上训练的。

相关的对策是:重新在微博上训练相关的工具,而这又需要标注大量数据或书写大量规则,这时可考虑利用主动学习或办监督学习的方法,利用现有的非微博上的工具和已标注数据,学习一个特定于微博的对应工具,而又减少或避免了数据标注或规则书写的高昂代价。

4　系统架构

我们介绍一种语义微博搜索的参考架构。如图 1 所示,该架构由 4 个部分构成:爬虫、单个微博处理模块、多个微博处理模块、索引。爬虫调用微博服务提供的开放接口获得微博,这一点跟一般搜索引擎通过 HTTP 协议盲目扫描互联网不同。索引模板对抽取到的信息点进行统一的多层次化的索引,多层次的含义是,不仅索引传统的关键字,还索引诸如命名实体、事件、情感等结构化的信息。索引本质是一个倒排表,其实现可参考开源软件(例如 Lucene 或 MySQL)。本节重点讨论微博处理模块,这些模块从单个或多个微博中抽取出信息点,是微博语义搜索的关键所在。

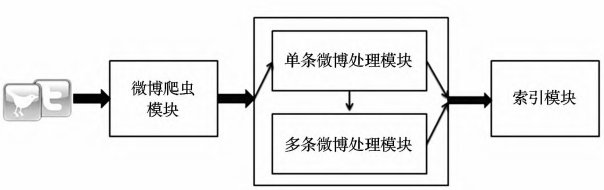

图 1　微博语义搜索的架构
Fig. 1　System architecture of microblog search

如图 2 所示,单个微博的处理和一系列机器学习及自然语言处理子模块相关。最重要的处理环节包括:分类、命名实体识别、语义角色标注和情感分析。分类的目的是将每个微博分到预定的类别中;命名实体识别是从微博中抽取出预先定义的各类实体,包括人物,机构,地点,时间,称谓等;语义角色标注是从微博中提取断言 - 参数结构化信息,它是其他类型信息(例如事件)抽取的基础;情感分析利用语义角色标注提供的信息,并综合其他方面的信息,从微博中提取情感的持有者、针对对象以及极性(正面、中性或负面);情感分析技术为商业智能提供了关键支持。

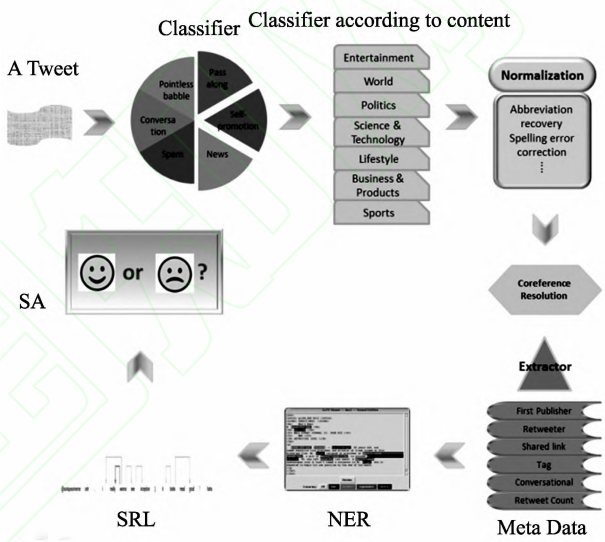

图 2　单个微博处理流程
Fig. 2　Processing pipeline for one microblog

多个微博处理模块构建在单微博处理模块基础之上,它负责从最近的微博集合中发现一些对用户有用的统计信息,如图 3 所示。微博不是孤立存在的,而是通过用户间的跟随关系、推荐关系关联到一起,形成各种类型的微博图。统计信息就是在微博图上进行的。重要的统计信息包括:最近热门话题、最近热门微博、最近热门人物、最近热门链接、最近热门博主、最近热门事件、最近热门观点等。此外,还从用户群中挖掘具有相似兴趣的用户社区,以及基于此之上的热门社区。这些统计信息从多个维度为用户提供最近微博的全貌。所有挖掘的统计信息,也按照分类微博所用的体系进行分类,这样允许用户以分类导航的方式浏览博客以及相关的统计信息。热门度可考虑两方面信息:活跃度或者出现频度(例如,热门人物一般是在微博中被讨论比较多的人物)和影响度或说可能被大家关注的程度(例如,热门博主往往发微博较多并且有较多的追随者)。

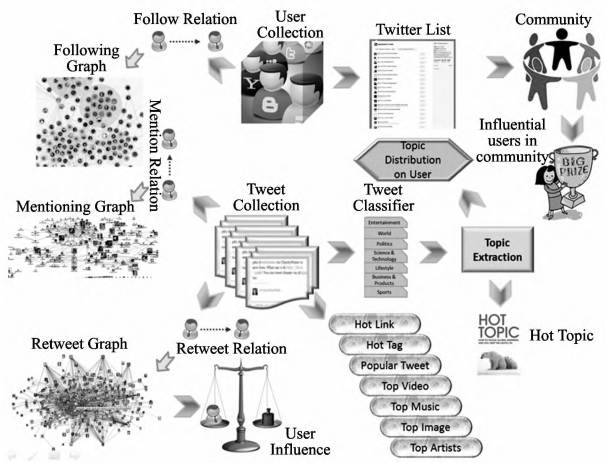

图3 多个微博处理流程

Fig. 3 Processing pipeline for multiple microblogs

5 关键技术

微博语义搜索是一组关键技术的集合。作为一种新的媒体形态,微博有和传统网页资源不同的特性,例如,微博短且噪音大,单个的信息量有限。面向微博的语义分析技术因为要和微博的这种特点匹配而具有了一些特殊性。作为一个样例,我们将考察语义角色标注模块的实现。该模块的实现对其他模块,例如情感分析、微博分类等的实现具有一般性的借鉴意义。

语义角色标注[2]是自然语言处理领域的一个经典任务。它从输入的文本中提取出断言-参数对。一般而言,断言由一个动词表示,参数描述动词的一个属性,例如施动者、受动者、动词指代的动作发生的时间、地点等信息。图4是一个语义角色标注的例子,A0、A1、AM-TMP、AM-LOC等表示角色名,前两者为核心角色,分别表示动词的施动者、受动者;后两者为辅助角色,分别表示动词的时间和地点参数。P表示断言,本例中为动词"sold"。

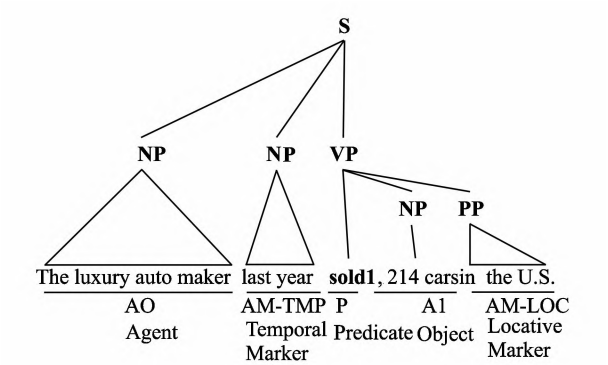

图4 语义角色标注的一个例子

Fig. 4 An illustrative example of semantic role labeling

传统的语义角色标注主要研究书写规范的文体,例如新闻。已经有一大批系统,例如基于流水线的方法,基于序列标记的方法,基于马尔科夫逻辑网的方法,在规范文本上取得了非常好的结果。但当这些系统应用到微博时,其性能明显下降[2]。这就要求一个专门针对微博的语义角色标注系统,而构建这样的一个专门系统,又需要大量的标注数据或者请专家编写大量规则,无论哪个方案,都代价昂贵。

一个可行的方案是:利用已有的语义角色标注资源,来自动标注一些微博,作为训练数据。具体而言,有一类微博是报道新闻事件的,而传统的新闻也报道了类似的事件。这样,可以先把报道同一事件的微博和新闻句子聚类到一起,然后用现有的语义角色标注系统对聚类中的新闻进行标注,接下来通过词对齐技术将新闻句子和微博中词分别对应起来,最后把新闻句子上的断言-参数结构根据词对齐的结果传递到微博上去。该过程如图5所示,其中,第一句为新闻句子,第二句是微博。新闻句子用现有的语义角色标注系统进行标注,微薄的语义角色信息则是通过词对齐由第一句映射而来。

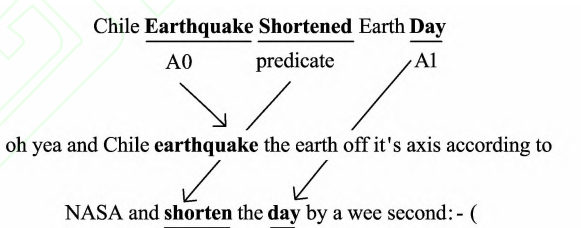

图5 利用聚类和词对齐对微博进行语义角色标注,以获得训练样本

Fig. 5 Obtaining training data using clustering and word alignment

有了标注语料后,就可以训练出统计模型,例如线性条件随机场模型。这种方法训练出来的语义角色标注系统的性能超过了现有的最好的语义角色标注系统[2]。尽管如此,有限的训练样本以及样本中的噪音仍然是影响性能的主要因素。

通过如下几个措施可进一步改进现有的语义角色标注系统。首先,可以利用主动学习的技术,与人互动,挑出那些最有价值的样本进行人工标注;其次,训练模型时,除了考虑传统特征外,还考虑微博特有的一些特征,例如链接、标签等;最后,可借鉴半监督学习的想法,把系统自己标注的一些高质量的样本加入到训练集,并用增强的训练集迭代训练模型。

语义角色标注模块的实现对其他关键模块的实现具有一定的启发意义。例如,在实现其他模块时,可考虑利用现有的资源,通过领域适配,获得一批训

练样本,尽量减少人工标注的代价;可综合考虑多个特征,以减少噪音对性能的影响;还可考虑利用微博本身之外的背景知识(例如相关微博)来增强上下文,以获得更丰富、更鲁棒的特征。

6　总结与展望

本文提出构建基于语义分析的微博搜索以克服现有微博搜索的不足。微博语义搜索从大量的微博中提取出信息点,以搜索和分类浏览的方式允许用户快捷地访问这些信息点。不同于现有的微博搜索,微博语义搜索对单个和一组微博做了深度的语义分析,提供了超越关键字搜索的高级搜索(例如搜索事件和搜索观点)和导航功能,并在一定程度上支持商业智能。本文分析了语义微博搜索的主要挑战和对策,并介绍了其参考实现和相关的关键技术。特别是,本文以语义角色标注为例,讨论了面向微博的语义分析技术如何克服微博本身特点带来的挑战。我们已经构造一个针对英文的微博语义搜索原型系统,下一步将扩展这一原型系统扩展以支持中文、日文等其他语言。

参考文献:

[1] 周明,刘晓华,蒋龙,等. 语义分析和搜索教程[R]. 北京:微软亚洲研究院,2010.

[2] LIU Xiaohua, ZHOU Ming, LI Kuan. SRL for news tweets[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010.

[3] 曹国伟. 新浪微博注册用户数已经突破 1.4 亿[EB/OL]. (2011-05-13)[2011-06-09]. http://news.ccidnet.com/art/3709/20110513/2389053_1.html.

[4] Jameson Berkow. FP tech desk: twitter now fastest-growing search engine[EB/OL]. (2010-07-08)[2011-05-24]. http://business.financialpost.com/2010/07/07/fp-tech-desk-twitter-now-fastest-growing-search-engine/

[5] Kelly Ryan. Twitter study reveals interesting results about usage[R]. San Antonio, Texas: Pear Analytics, 2009.

(编辑:许力琴)