

文章编号:1001-9081(2015)04-1013-04

doi:10.11772/j.issn.1001-9081.2015.04.1013

基于自监督学习的维基百科家庭关系抽取

朱苏阳^{1,2*}, 惠浩添^{1,2}, 钱龙华^{1,2}, 张民^{1,2}

(1. 苏州大学 自然语言处理实验室, 江苏 苏州 215006; 2. 苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

(* 通信作者电子邮箱 20124227049@suda.edu.cn)

摘要:传统有监督的关系抽取方法需要大量人工标注的训练语料,而半监督方法则召回率较低,对此提出了一种基于自监督学习来抽取人物家庭关系的方法。该方法首先将中文维基百科的半结构化信息——家庭关系三元组映射到自由文本中,从而自动生成已标注的训练语料;然后,使用基于特征的关系抽取方法从中文维基百科的文本中获取人物间的关系。在一个人工标注的家庭关系网络测试集上的实验结果表明,该方法优于自举方法,其F1指数达到77%,说明自监督学习可以较为有效地抽取人物家庭关系。

关键词:自监督学习;维基百科;半结构化信息;关系抽取

中图分类号: TP391 **文献标志码:**A

Family relation extraction from Wikipedia by self-supervised learning

ZHU Suyang^{1,2*}, HUI Haotian^{1,2}, QIAN Longhua^{1,2}, ZHANG Min^{1,2}

(1. Natural Language Processing Laboratory, Soochow University, Suzhou Jiangsu 215006, China;

2. School of Computer Science and Technology, Soochow University, Suzhou Jiangsu 215006, China)

Abstract: Traditional supervised relation extraction demands a large scale of manually annotated training data while semi-supervised learning suffers from low recall. A self-supervised learning based approach was proposed to extract personal family relationships. First, semi-structured information (family relation triples) was mapped to the free text in Chinese Wikipedia to automatically generate annotated training data. Then family relations between person entities were extracted from Wikipedia text with feature-based relation extraction method. The experimental results on a manually annotated test family network show that this method outperforms Bootstrapping with F1-measure of 77%, implying that self-supervised learning can effectively extract personal family relationships.

Key words: self-supervised learning; Wikipedia; semi-structured information; relation extraction

0 引言

家庭关系抽取(Family Relation Extraction, FRE)是指从文本中抽取两个人物之间的家庭关系,从而为社会关系网络的构建提供基础。家庭关系抽取是命名实体间语义关系抽取(简称关系抽取)的一种特殊情况,即两个实体均为人物。根据所使用语料库数量的不同,基于机器学习的关系抽取方法可分为有监督学习^[1-3]、半监督学习^[4]和无监督学习^[5]等。有监督学习方法也被用于从文本中抽取社会关系网络^[6-7]。其特点是标注语料的规模、质量和领域决定了关系抽取的性能,但是对于网络上大量的自然语言文本,手工标注大量的高质量训练语料显然不太现实。

半监督学习方法从小规模的标注语料库中出发,不断从大规模未标注语料中自动挖掘可靠性高的标注实例充实到训练语料中,从而构建出具有一定规模的训练语料库。半监督学习中的代表性算法为自举学习(Bootstrapping),它只需要极少量的种子集。自举算法被应用于中文家庭关系网络的构建^[8],从大规模未标注文本中挖掘出特定的家庭关系实例,

其优点是准确率较高,但召回率较低,且易产生语义漂移^[9]。

近几年涌现出的一种新的学习方法——自监督学习(Self-Supervised Learning, SSL),其训练语料库不是由人工标注,而是利用结构化数据库^[10-11]或半结构化信息^[12-14]中的关系实例自动映射到自然语言文本中,从而产生较为可靠的训练语料。该方法被广泛应用于开放型信息抽取(Open Information Extraction, OIE)中,即判断两个实体间是否存在语义关系,而不关心这一关系是否属于预先指定的关系类型。自监督学习避免了有监督学习中标注大量语料库所需要的人力,也缓解了半监督学习中所存在的问题。

有鉴于此,本文提出将自监督学习方法应用于家庭关系抽取。首先,将维基百科的半结构化数据中的人物家庭关系实例对应到维基百科中的文本中,并自动生成标注的训练语料;再利用基于特征向量的关系抽取方法从维基百科的文本中抽取人物家庭关系。在一个小规模的家庭关系网络上的实验结果表明,借助于半结构化信息,自监督学习方法能够较为有效地从中文维基百科中抽取特定类型的人物家庭关系。

收稿日期:2014-10-27;修回日期:2015-01-05。

基金项目:国家自然科学基金资助项目(61373096, 90920004);江苏省高校自然科学研究重大项目(11KJA520003)。

作者简介:朱苏阳(1989-),男,江苏苏州人,硕士研究生,主要研究方向:信息抽取;惠浩添(1991-),男,江苏徐州人,硕士研究生,主要研究方向:信息抽取;钱龙华(1966-),男,江苏苏州人,副教授,CCF会员,主要研究方向:自然语言处理;张民(1970-),男,黑龙江哈尔滨人,教授,博士生导师,CCF会员,主要研究方向:机器翻译。

1 基于自监督的家庭关系抽取

图1为基于自监督学习的家庭关系抽取的流程。首先从维基百科的页面中筛选出人物页面,再从其中提取出文本和家庭关系三元组;然后将关系三元组自动映射到文本中,从而获得训练语料;接着训练出分类模型;最后应用分类模型从人物页面的文本中抽取新的家庭关系实例。

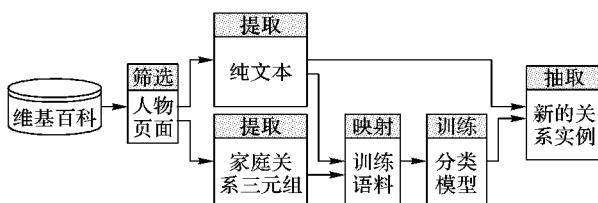


图1 基于自监督学习的家庭关系抽取的流程

1.1 维基百科人物实体列表的获取

维基百科中每一个页面 a 可定义为一个五元组^[15]: $a = (title, text, ib, tp, C)$,即维基百科中的一个页面由标题(Title)、正文(Text)、信息盒(Infobox)、模板(Template)、类型(Category)等5个部分构成。本文采用启发式规则的方法,根据一个页面所属的类型来判断该页面是否属于人物类型。举例来说,页面“李白”属于以下几个类型:701年出生、762年逝世、唐朝诗人、唐朝作家、平凉人、李姓、水神、人物神。通过判断页面李白属于XX年出生、XX年逝世这两个类型就可以认定该页面属于人物类型,且该页面的标题即是所对应人物的一个正式名称。对于人物实体的多名问题,可以通过维基百科的重定向列表获得这些人物实体的其他名称,如苏轼(正式名称)有苏东坡、苏和仲等其他名称。

1.2 家庭关系的定义

信息盒是维基百科中的半结构化信息,其中的每一个表项都可以视为一个三元组 $ib = (title, attr, value)$,即(标题,属性名,属性值)。例如(刘禅,父亲,刘备)这样一个三元组是文章 $a_{刘禅}$ 中的一个信息盒表项,它表示了人物实体刘禅与刘备之间的亲子关系。

本文借助哈尔滨工业大学的同义词词林扩展版(<http://www.ltp-cloud.com/download/>),将其中Ah大类(即亲人/亲戚/眷属)中的家庭关系名称作为关键词,将其与信息盒中的属性名称作部分匹配,从而获取信息盒中表示人物家庭关系的属性名称。对所得的结果作人工筛选,剔除部分较为口语化的关系名称,最终获得信息盒中的家庭关系名称共133个。本文将这些关系分为6类,其中部分关系的分类体系如图2所示。

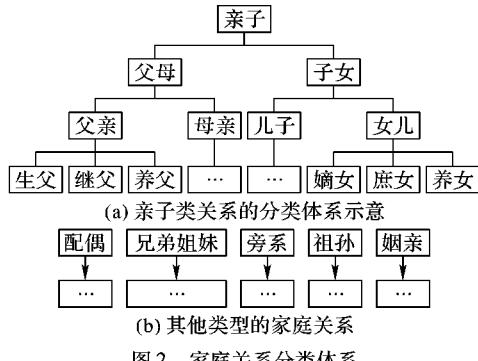


图2 家庭关系分类体系

具体为:

- 1) 亲子关系(Parents-Children,P&C);
- 2) 配偶关系(Spouse,SPO);
- 3) 兄弟姐妹关系(Sibling,SIB);
- 4) 旁系亲属关系(Collateral Relatives,COL);
- 5) 祖孙关系(Grandparent and Grandchild,GP&C);
- 6) 姻亲关系(Affinity,AFF)。

1.3 训练数据集的生成

自监督方法将所得的家庭关系三元组按照一定规则映射到自由文本中来获取正例,再通过其他规则从除正例之外的剩余数据中获取负例。其流程如下:

- 1) 全体句子集合 S_t 的构建。

在得到了维基百科中所有的人物页面后,本文要构造一个包含这些人物实体对的句子集合 S_t 。具体过程是:首先对维基百科中的人物页面文本使用中文分词工具ICTCLAS进行分词,然后根据上述人物实体列表对句子通过字符串匹配的方法自动标注出人物实体,再对所得句子 s 采用以下规则进行处理。

①若句子 s 包含 k ($k \geq 2$)个不同人物实体,则取这些人物实体的两两组合,将 $C(k,2)$ 个句子 s 加入到集合 S_t 。由于这些句子包含不同的人物实体对,因此视为不同的关系实例。

②若句子 s 仅包含一个人物实体 e ,并且 e 不是句子 s 所在页面的人物实体,则将句子 s 加入到集合 S_t 。这一规则主要考虑到在文章 a 中可能有部分句子存在主语缺省现象,例如文章 $a_{长孙无忌}$ 中的句子“祖父长孙无忌,北周开府仪同三司,平原县侯。”,该句在句首省略了所在页面的人物实体“长孙无忌”。

构建完成的集合 S_t 包含576504个关系实例,它可以分为三个子集:已知训练中所用到的正例集合 S_p 、已知负例集合 S_n ,以及未知类型的集合 S_u 。

- 2) 正例集合 S_p 的构建。

自监督学习通过将维基页面中已有的半结构化信息(即信息盒)映射到文本中的方法来获取正例集合 S_p 。根据已定义的家庭关系,从人物页面的信息盒中共获取24688条家庭关系三元组 $r(e_1, e_2)$ (其中的人物实体 e_1 的名称即为该三元组所在页面的标题),所有这些三元组中的人物实体对 (e_1, e_2) 构成一个具有家庭关系的实体对集合 E_{tuple} (该集合考虑两个实体的先后顺序),再将这些三元组映射到其所属页面的正文句子中。具体方法是:对于三元组 $r(e_1, e_2)$ 成立且句子 s 所在页面的标题所指人物实体即为 e_1 的句子 $s \in S_t$,根据以下规则从全集 S_t 中构建正例集合 S_p :

①若 e_1, e_2 同时出现在句子 s 中,则认为句子 s 是家庭关系 r 的一个正例实例,将句子 s 加入到集合 S_p 。

②若表示家庭关系 r 的关键词与 e_2 同时出现在句子 s 中,而 e_1 或 e_1 的其他名称不出现在句子 s 中,则认为句子 s 是家庭关系 r 的一个正例实例,将句子 s 加入到集合 S_p 。这一规则与上一条相类似,只不过是句子中存在主语缺省现象而已。

构建完成的集合 S_p 有9290个实例,集合 S_p 中的正例所属页面的人物构成人物集合 E_{title} 。各类家庭关系的正例数见表1。从表1可看出:所得的正例中亲子关系所占比例最高(约65%),其次是配偶关系(约26%)和兄弟姐妹关系(约7%),而旁系亲属、祖孙、姻亲关系所占比例极少(均小于1.5%)。

需要说明的是,使用上述方法产生的正例集合中会包含一定的噪声,也可能丢失真正的正例,而前者比后者产生的问题更严重。通过随机抽样分析发现,正例集合中包含约20%的噪声。

表1 6类家庭关系的正例数及所占比例

类型	数量	所占百分比/%
P&C	6010	64.7
SPO	2415	26.0
SIB	640	6.9
COL	35	0.4
GP&C	65	0.7
AFF	125	1.3
总计	9290	100.0

3) 负例集合 S_n 的构建。

对于自监督学习来说,负例的生成也是一个相当关键的问题。当从全集 S_t 中提取出正例集合 S_p 后,剩余的实例并非都是负例。本文产生负例的基本思想是:当句子 s 中没有出现家庭关系关键词时才把它作为一个潜在的负例。具体而言,对不包含家庭关系关键词的句子 $s \in S_t - S_p$ 应用以下规则:

①若句子 s 中仅包含一个人物实体 e ,则将句子 s 加入到集合 S_n 。

②若句子 s 中包含两个人物实体 e_1 和 e_2 ,且 $e_1 \in E_{\text{title}}$ 或 $e_2 \in E_{\text{title}}$,同时 $(e_1, e_2) \notin E_{\text{tuple}}$ 或 $(e_2, e_1) \notin E_{\text{tuple}}$,则将句子 s 加入到集合 S_n 。

在句子中不出现家庭关系关键词的前提下,第一条规则的含义是指句子中的缺省主语和唯一的人物实体构成一个负例;第二条规则的含义是指正例集合 S_p 中出现的某一人物和该集合之外的人物不构成任何家庭关系,当然这个假设并不总是成立,因而负例集合中也会包含一定的噪声(负例噪声比例约1.9%)。构建完成的集合 S_n 包含140717条负例,而剩余的未知类型实例集合 S_u 即为 $S_t - (S_p \cup S_n)$ 。

1.4 特征选择

本文使用的特征均为基于人物实体对的词汇特征与统计特征,具体如下(假设在句中人物实体 e_1 位于 e_2 之前):

- 1) 人物实体 e_1 与 e_2 间的词距离;
- 2) 人物实体 e_1 与 e_2 间的子句距离;
- 3) 人物实体 e_1 与家庭关系关键词 r 的词距离;
- 4) 人物实体 e_2 与家庭关系关键词 r 的词距离;
- 5) 人物实体 e_1 的前5个词;
- 6) 人物实体 e_2 的后5个词;
- 7) 人物实体 e_1 与 e_2 之间的词。

2 实验与分析

2.1 实验设置

本文使用构建好的正负例集合作为训练语料,分别采用内部测试与外部测试两种测试方法。分类器使用基于支持向量机(Support Vector Machine, SVM)分类器 SVM-Light 的多元分类器 SVM-Multiclass。评估标准采用常用的准确率 P 、召回率 R 以及调和平均值 $F1$ 指数。在核函数的选取上,由于样本数与特征维度均很大,采用缺省的线性核^[16]。

2.2 内部测试

内部测试的数据集为自动产生的正例集合 S_p 和负例集

合 S_n ,采用5倍交叉验证的策略检验分类模型的泛化能力。由于正负例比例严重不平衡(正:负约为1:15),本文对训练集中的负例采用随机欠采样的方法,而测试集中的正负比例则采用其原始值1:15。实验结果表明,当训练集中的正负例比例为1:3时,内部测试的性能最好。表2给出了平均 $F1$ 值最高的情况下(正:负=1:3)各关系大类的性能。

表2 自监督方法在内部测试中平均 $F1$ 值最高时的结果 %

关系类型	准确率	召回率	$F1$
P&C	78.0	78.8	78.4
SPO	53.4	51.3	52.3
SIB	55.4	31.1	39.8
COL	0.0	0.0	—
GP&C	17.7	16.9	17.3
AFF	3.8	4.0	3.9
平均	68.3	66.6	67.2

从表2可看出内部测试的平均 $F1$ 指数为67.2,总体性能并不高。与表1对照可发现,关系大类的性能与实例数量密切相关,实例数量越多的关系性能越好。进一步的分析表明,自动产生的数据集中的噪声和测试集中正负例的不平衡现象,是导致内部测试性能不高的两个主要原因。通过抽样检测发现,本文构造的正负例集合中的噪声比例分别为20%和2%。由于正负例比例约为1:15,因此一个理想的分类模型在测试集上的准确率也只有 $1 \times (1 - 0.2) / (1 \times (1 - 0.2) + 15 \times 0.02) \times 100\% \approx 73\%$,召回率为 $(1 - 0.2) \times 100\% = 80\%$, $F1$ 指数约为77%。

既然上述测试集中存在着噪声,那么其性能也并不代表该分类模型在真实数据集上的性能,因此需要外部测试来评估分类模型的真实泛化能力。

2.3 外部测试

外部测试是指在自动产生的正例集合和负例集合以外的数据集上测试其泛化能力。本文手工标注了一个小规模的家庭关系网络以检验分类模型的抽取性能。该网络以中文维基百科中的【康熙帝】、【雍正帝】、【乾隆帝】三个人物为中心节点,从相应的页面文本中标注出其中的人物家庭关系(其中包括信息盒中已有的关系),再从这些家庭关系延伸到其他人物的页面文本,最终构成一个共涉及5代人的家庭关系网络。

该家庭关系网络共包含人物156个,家庭关系443对,其中在文本中出现的关系共267对,有176对关系只出现在信息盒中。本文以前者作为测试集,从相应的页面文本中抽取包含这些人物对的句子作为测试实例,最终得到445个测试实例。训练集中的正负例比例采用内部测试中 $F1$ 最高的1:3,并且从中去除包含测试实例的句子。由于测试集中部分关系的实例数较少,可能导致实验结果出现较大的抖动,因此训练时对负例进行5次随机欠采样,并对实验结果取平均值。实验结果如表3所示。

本文实现了典型的自举算法 Espresso 系统^[17]作为基准系统与自监督方法进行比较,并评估该基准系统在家庭关系网络上的性能。基准系统中每类关系所选用的种子均来自信息盒中已知的关系实例,每种关系类型的种子数为4~10个。在每轮迭代中选取置信度最高的前3~5的实例与模式。基

准系统先在全体中文维基语料上挖掘出家庭关系实例集合,然后再取所得实例与测试集中的家庭关系网络中实例的交集作为测试所得结果。测试性能如表 4 所示。

表 3 自监督方法在外部测试中的结果 %

关系类型	网络中实例数	准确率	召回率	F1
P&C	189	75.6	88.4	81.5
SPO	28	93.8	67.9	78.8
SIB	31	90.0	58.1	70.6
COL	4	0.0	0.0	—
GP&C	15	75.0	33.3	46.1
AFF	0	—	—	—
平均	—	78.0	76.6	77.3

表 4 基准系统在网络上的性能 %

关系类型	网络中实例数	准确率	召回率	F1
P&C	189	88.9	6.7	12.5
SPO	28	0.0	0.0	—
B&S	31	0.0	0.0	—
COL	4	0.0	0.0	—
GP&C	15	16.7	7.1	10.0
AFF	0	—	—	—
平均	—	60.2	5.0	9.2

从表 3 可看出:外部测试的性能明显高于内部测试,其平均 F1 指数达到 77.3%,特别是 SPO 和 SIB 关系,性能有显著提高,F1 指数均超过 70.0%。而从表 4 中看出,自举系统虽然在 P&C 关系上准确率很高,但整体召回率很低,因而并不适合于文本的深度挖掘。这是由于自举系统仅选取语料库中出现频度和准确率两者都较高的模式,而这些模式的覆盖度较低,导致整体性能下降,因此自举方法更适合于从大规模语料中发现高频率的关系实例。

2.4 错误分析

外部测试的性能离预期还有一定的差距。通过对分类结果的错误分析发现,外部测试中出现的错误主要与特征选取、篇章结构和语料库噪声有关,具体可分为以下三类错误:

1) 词语特征过于泛化或稀疏。前者是指某些常用特征词语总是和特定的关系类型相关联,而后者指某些生僻词语在训练语料中没有出现。例句“【承庆】是【胤禔】的同母哥哥”中分类器根据“母”一词就错分为父子关系,而“【惠妃】,【康熙帝】的妃嫔。”中生僻词“妃嫔”导致它被错分为负例。该类错误约占全体错误的 80%。

2) 成分缺省。为了考虑篇章的衔接性,句子中的各种成分经常被省略。例如句子“顺治七年,叔父摄政王【多尔袞】去世,很快十四岁的【顺治帝】开始亲政。”中的第二个子句前省略了“顺治帝”,给关系抽取带来了困难。该类错误约占全体错误的 7%。

3) 无法明显归类的错误,可能由于训练集噪声而被错误分类。该类错误约占全体错误的 13%。

3 结语

本文使用自监督学习方法,利用中文维基百科上的半结构化信息,从维基百科的自然文本中自动获取训练数据,并在一个人工标注的家庭关系网络上测试了分类模型的关系抽取

性能。实验结果表明这一方法在中文维基百科上进行家庭关系网络的构建时行之有效,尤其是对于常用的家庭关系类型。然而,由于训练数据的噪声与特征选择的原因,对于训练语料较少的关系类型,抽取性能还不够理想。下一步的工作将会从减少训练集噪声、改进抽取特征两个方面展开,以提高家庭关系的整体抽取性能。

参考文献:

- [1] ZHOU G, ZHANG M. Extracting relation information from text documents by exploring various types of knowledge [J]. Information Processing and Management, 2007, 43(4): 969 – 982.
- [2] ZHOU G, QIAN L, FAN J. Tree kernel-based semantic relation extraction with rich syntactic and semantic information [J]. Information Science, 2010, 180(8): 1313 – 1325.
- [3] OH J-H, UCHIMOTO K, TORISAWA K. Bilingual co-training for monolingual hyponymy-relation acquisition [C]// ACL 2009: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language. Stroudsburg: Association for Computational Linguistics, 2009: 432 – 440.
- [4] ZHU Z. Weakly-supervised relation classification for information extraction [C]// Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2004: 581 – 588.
- [5] ZHANG M, SU J, WANG D, et al. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering [C]// Proceedings of the Second International Joint Conference on Natural Language Processing. Berlin: Springer-Verlag, 2005: 378 – 389.
- [6] JING H, KAMBHATLA N, ROUKOS S. Extracting social networks and biographical facts from conversational speech transcripts [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2007: 1040 – 1047.
- [7] AGARWAL A, CORVALAN A, JENSEN J, et al. Social network analysis of Alice in Wonderland [C]// Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. Stroudsburg: Association for Computational Linguistics, 2012: 88 – 96.
- [8] GU J, HU Y, QIAN L, et al. Research on building family networks based on bootstrapping and coreference resolution [C]// NLPCC 2013: Proceedings of the Second CCF Conference on Natural Language Processing and Chinese Computing, CCIS 400. Berlin: Springer-Verlag, 2013: 200 – 211.
- [9] KOMACHI M, KUDO T, SHIMBO M, et al. Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms [C]// EMNLP 2008: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2008: 1011 – 1020.
- [10] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data [C]// ACL 2009: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009: 1003 – 1011.
- [11] RIEDEL S, YAO L, McCALLUM A. Modeling relations and their mentions without labeled text [C]// Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2010: 148 – 163.

(下转第 1020 页)

的相关性数据如图 3 所示。

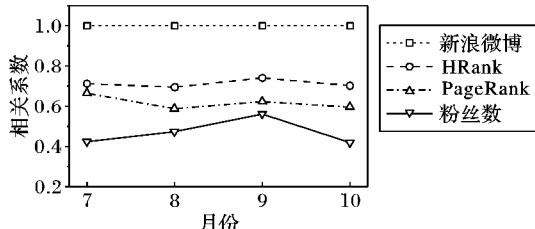


图 3 各影响力模型与新浪官方影响力排名相关性

由图 3 可看出:粉丝数与用户影响力的的相关性不是很强,同样数据集下相对 PageRank,HRank 用户影响力模型与新浪用户影响力官方排名更为接近。

4 结语

用户影响力评价作为微博平台数据挖掘领域的研究热点,通过发现微博消息传播网络中的强力节点,可为相关管理者依据强力用户的传播机制进行广告投放、舆情管控等操作提供依据。本文对微博用户影响力的研究现状进行了分析,以 Hadoop 集群为主要计算平台研究了基于用户粉丝和微博转发数的 H 指数对用户影响力的评价效果,并提出了 HRank 用户影响力评价模型,该模型既考虑了用户间的静态关系特征,又综合了用户的动态行为特性,从用户的粉丝质量和用户的微博质量来考量用户的影响力,并且很大程度降低了僵尸粉和垃圾微博对评价结果的影响。通过与 PageRank 用户影响力模型,以及新浪微博现有的用户影响力模型进行对比,发现 HRank 用户影响力评价取得了与实际最接近的效果。在下一步的研究工作中,将继续结合其他实际应用领域,深入探讨算法的有效性和实用性。

参考文献:

- [1] WU K, JI X, GUO J, et al. Influence maximization algorithm for micro-blog network [J]. Journal of Computer Applications, 2013, 33(8): 2091–2094. (吴凯, 季新生, 郭进时, 等. 基于微博网络的影响力最大化算法[J]. 计算机应用, 2013, 33(8): 2091–2094.)
- [2] QI C, CHEN H, YU H. Method of evaluating micro-blog users' influence based on comprehensive analysis of user behavior [J]. Application Research of Computers, 2013, 31(7): 2004–2007. (齐超, 陈鸿昶, 于洪涛. 基于用户行为综合分析的微博用户影响力评价方法[J]. 计算机应用研究, 2013, 31(7): 2004–2007.)
- [3] CAO J, WU J, SHI W, et al. Sina microblog information diffusion analysis and prediction [J]. Chinese Journal of Computers, 2014, 37(4): 779–790. (曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779–790.)
- [4] CHA M, HADDADI H, BENEVENUTO F, et al. Measuring user influence in Twitter: the million follower fallacy [C]// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2010: 10–17.
- [5] CHEN H. Microblog user ranking research based on Hadoop [D]. Shanghai: East China University of Science and Technology, 2014. (陈浩. 基于 Hadoop 的微博用户影响力排名算法研究[D]. 上海: 华东理工大学, 2014.)
- [6] KANG S. The evaluation of the social network's nodes influence based on users' behavior [D]. Beijing: Beijing University of Posts and Telecommunications, 2011. (康书龙. 基于用户行为及关系的社交网络节点影响力评价[D]. 北京: 北京邮电大学, 2011.)
- [7] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: quantifying influence on Twitter [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 65–74.
- [8] HIRSCH J E. An index to quantify an individual's scientific research output [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569–16572.
- [9] YU L, HU C, SU L. Micro-blogs data collection based on MapReduce [J]. Computer Science, 2012, 39(11A): 143–145. (于留宝, 胡长军, 苏林哈. 基于 MapReduce 的微博文本采集平台[J]. 计算机科学, 2012, 39(11A): 143–145.)
- [10] SHU Y, XIANG Y, ZHANG Q, et al. Research on MapReduce algorithm of micro blog ranking [J]. Computer Technology and Development, 2013, 23(2): 73–76. (舒琰, 向阳, 张琪, 等. 基于 PageRank 的微博排名 MapReduce 算法研究[J]. 计算机技术与发展, 2013, 23(2): 73–76.)
- [11] PING Y, XIANG Y, ZHANG B, et al. Implementation of parallel PageRank algorithm based on MapReduce [J]. Computer Engineering, 2014, 40(2): 31–34. (平宇, 向阳, 张波, 等. 基于 MapReduce 的并行 PageRank 算法实现[J]. 计算机工程, 2014, 40(2): 31–34.)
- [12] ZHANG Y. Research on information dissemination and opinion evolution in the social networking services [D]. Beijing: Beijing Jiaotong University, 2012. (张彦超. 社交网络服务中信息传播模式与舆论演进过程研究[D]. 北京: 北京交通大学, 2012.)

(上接第 1016 页)

- [12] HOFFMANN R, ZHANG C, LING X, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C]// HLT 2011: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 541–550.
- [13] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance multi-label learning for relation extraction [C]// EMNLP-CoNLL 2012: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2012: 455–465.
- [14] WU F, WELD D. Open information extraction using Wikipedia [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 118–127.
- [15] WANG Z, LI Z, LI J, et al. Transfer learning based cross-lingual knowledge extraction for Wikipedia [C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 641–650.
- [16] HSU C, CHANG C, LIN C. A practical guide to support vector classification [R]. Taipei: University of National Taiwan, 2003.
- [17] PANTEL P, PENNACCHIOTTI M. Espresso: leveraging generic patterns for automatically harvesting semantic relations [C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006: 113–120.