基于自然语言处理的弱监督知识获取系统的实现

田 东[1] 张西宁[2]
(1.西安邮电大学 计算机学院 西安 710121;2.陕西广电网络传媒(集团)股份有限公司技术部 西安 710061)

摘 要:知识获取多年来一直被认为是阻碍智能系统开发的瓶颈问题,尤其是互联网时代,大量的信息都以非结构化的文本形式存在。本文运用分布式计算思想设计了一个基于互联网大规模语料库的知识自动获取系统。采用弱监督条件下机器学习的方法对信息自动挖掘和获取,实现机器对知识的自动学习和挖掘、新词词典发现、实体关系模板提取、命名实体识别等功能。利用该系统分别对未登录新词发现和地名识别两种应用进行了实验,运用 N-gram 和互信息(PMI)方法分别取得了72.1%和87.28%的准确率。

关键词:自然语言处理;分布式计算;弱监督机器学习;知识获取
中图分类号: TN081 **文献标识码**:A **国家标准学科分类代码**: 520.2020

Implementation of weakly supervised learning knowledge acquisition system based on natural language processing

Tian Dong[1], Zhang Xining[2]
(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; 2. Technology Department, Shaanxi Broadcast & TV Network Intermediary (Group) Co. Ltd., Xi'an 710061, China)

Abstract: Knowledge acquisition has been considered as a bottleneck problem in the development of intelligent systems for many years. Especially in the Internet era, a large number of information exists in the form of unstructured text. This paper introduces a knowledge acquisition system for a large Web page corpus based on distributed computing. This system is designed for automatic information mining and acquisition by the weakly supervised learning method. Computers can realize the automatic learning and mining of knowledge, the discovery of new words dictionary, the extraction of entity relation template, the entity recognition and so on. We represent the N-gram model and pairwise mutual information methods for new words recognition and location name entity detection, and the experimental results show the precision are 72.1% and 87.28% respectively.

Keywords: natural language processing; distributed computing; weekly supervised learning; knowledge acquisition

1 引 言

随着互联网的迅猛发展和广泛普及,网络信息越来越丰富,新信息的产生和传播日益迅速。对于每天都在大量产生的新信息,以及新的语言运用方式,需要对其保持跟踪和更新,以适应当今知识获取的需要。

在信息挖掘的早期研究中,研究者主要是使用模板技术作为挖掘信息的主要手段。早期系统的主要缺陷是人工干预较强,对模板和语法的依赖太过严重;概率评估效果不足,概率评估更多是对语法分析的一种补充;信息挖掘方向的局限性太强且扩展性不佳。Oren Etzioni 和 Michael Cafarella 等人[1]于 2004 年设计完成了无监督的信息挖掘系统 KnowItAll,在英文信息挖掘方面取得了显著的成果。但由于中文结构复杂,语法分析与英文存在差别等原因,KnowItAll 对中文的支持性并不理想。

以 KnowItAll 思想为基础,在 Hadoop 平台[2]上运用 map reduce 分布式计算方式设计实现了一种基于中文自然语言处理的弱监督知识获取系统,实现机器对信息的自动学习和挖掘、模板提取、命名实体识别等功能,可以很好地满足中文知识获取的需要,适应了自然语言处理在大数据条件下的发展和应用。

收稿日期:2017-01

2 系统组成与处理流程

2.1 系统组成

通过对基于自然语言处理的弱监督信息挖掘系统的研究开发，实现了一个开放通用的信息挖掘平台，可以将系统设计分为4大模块，即数据抓取模块、自然语言处理模块、信息预处理模块和信息挖掘模块[3]。

2.1.1 数据抓取模块

数据的抓取方式，可以分为两类。1)对网络站点直接抓取，获取大批量的语料；2)使用基于特定模板规则的自动定制搜索，每一个搜索查询都是根据特定的模板相互关联的关键词组成的规则。

2.1.2 自然语言处理模块

中文自然语言处理模块采用了层叠自动机规则法和统计方法相结合的混合模式，对于文本实现了分词、词性标注、实体命名标注、地点归一化、时间归一化、实体关系识别，结果以XML的格式输出。自然语言处理模块架构如图1所示。

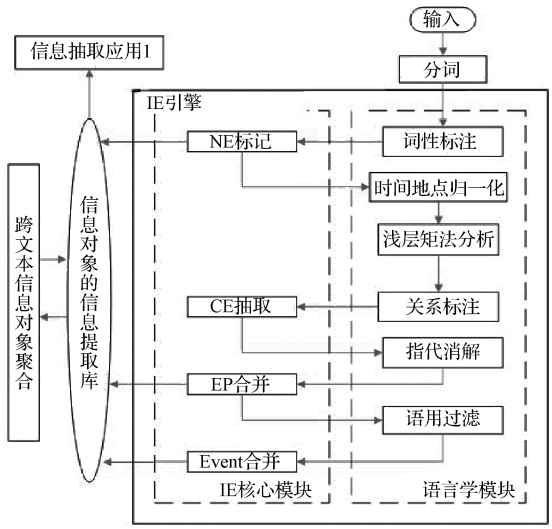

图1 自然语言处理模块架构

2.1.3 信息预处理模块

对信息的挖掘一般都是以词为单位，而自然语言处理生成的XML文件[4]是以一篇文章为单位的，并不能直接被系统用来挖掘信息。对于不同的信息挖掘需求，需要对XML文件进行必要的预处理操作，生成固定的数据结构，以适应信息挖掘的需要。本系统中主要采用以下两种数据结构：

1)window-size结构

自然语言处理的统计规律表明，相关词汇的出现总是遵循一定的共生关系，即它们共同出现的概率远远大于与其它词汇搭配出现的概率。自然语言处理的词典编写经验表明，实际语料中的大多数组合词汇关系，在句子中的距离一般都不会超过5个词[5]。因此，对于一个给定的名词作为中心词，检查中心词附近的词，范围取在中心词相邻的5个词之内，组成如下结构：

(中心词，相邻词，两词之间的距离)

这里“两词之间的距离”为中心词与相邻词之间间隔了几个词，相邻词位于中心词之后为正数，位于中心词之前为负数。

例如，“我是中国人”，经过自然语言处理分词后为“我是 中国人”。可以提取window-size数据结构为：(我，是，0)、(我，中国人，1)、(中国人，我，-1)等。

2)三元组结构

在实际的自然语言处理过程中，有些时候更加关注词语在语料之间的语法关系，通过语法关系更容易挖掘和构建特定的模板。因此我们定义了一种三元组关系。一个三元组由输入语句中的两个词和它们在句子之间的语法关系构成，其构成形式为：

(输入词1，输入词2，两词之间的关系)

例如，对于句子“我是中国人”，可以提取的三元组是：(我，中国人，指代关系)、(我，是，主谓关系)等。

2.1.4 信息挖掘模块

自然语言处理过程中，经常会看到“某某生于某地”，“某某担任某职”等语料，且它们的词性、格式等性质都在某些方面符合一定的统计规律。因此，可以通过统计计算，从语料中提取某些固定的模板，并将其应用到实际的自然语言处理中去。例如，对于“织田信长出生于日本尾张那古野城(今爱知县)”，使用模板“某人(NePer) 出生于某地(NeLoc)”，可以判定织田信长的词性是一个人物的概率较大，而尾张那古野城(今爱知县)是一个地名的概率较高。

2.2 系统处理流程

系统处理流程主要分为以下几个步骤：

S1：使用网络抓取模块抓取网络数据；

S2：使用中文语言处理模块对语料进行分析，生成XML文件；

S3：使用信息预处理模块对数据进行预处理；

S4：加入诱导种子，并使用信息挖掘系统模块中的筛选程序对数据进行筛选；

S5：对筛选数据进行概率评估，抽取信息模板；

S6：使用抽取到的模板筛选数据，将数据加入到诱导种子队列中；

S7：返回步骤S4，循环执行，直到数据收敛或达到预定条件后停止；

S8：对实验结果进行人工分析。

3 弱监督机器学习方法

3.1 Bootstrapping技术

为了评估语料的概率和价值，需要反复且大量地对语料进行输入和训练，这对于少量输入数据来说较为容易，

但面对大量数据时就变得相当繁琐且不安全[6]。系统设计时，力求简单易用，并且实现一定程度的自动化管理，最终目标是实现完全自动化。这要求系统能够尽量减少人工操作，并对数据进行自动化的训练。为此，本文设计了一种引导技术来实现 Bootstrapping。

一方面，使用诱导种子从通用提取模式提取数据。在数据挖掘的开始阶段，首先使用诱导种子来分析数据，凡是满足诱导种子模板的数据都提取出来。例如，在城市地名发现实验中，使用“北京”、“天津”、“西安”等作为种子，并使用模板“NeLoc 市”作为过滤模板，只有包含种子并且符合模板的语料才会被提取出来。

另一方面，通过概率评估，自动生成鉴别模板。对诱导种子抽取的信息进行分词和词性标注，然后结合朴素贝叶斯公式[7]，使用式(1)进行模板的概率评估。

$$P(\varphi \mid f_1, f_2, \ldots, f_n) = \frac{P(\varphi)\prod_i P(f_i \mid \varphi)}{P(\varphi)\prod_i P(f_i \mid \varphi) + P(\neg\varphi)\prod_i P(f_i \mid \neg\varphi)} \quad (1)$$

$P(\varphi)$ 表示语料的数据，是一个根据语料计算出来的先验概率。$P(f_i \mid \varphi)$ 表示在 φ 出现时，f_i 出现的概率，是一个条件概率。$P(\varphi \mid f_1, f_2, \ldots, f_n)$ 表示 φ 在使用 f_1，$f_2, \ldots, f_n$ 为模板时的概率评估。通过寻找最大的概率，可以提取出相应的模板：

$$f_1, f_2, \ldots, f_n = \text{argMax}P(\varphi \mid f_1, f_2, \ldots, f_n)$$

3.2 概率评估能力

本系统中主要使用了两种评估方式进行概率评估。

1)点互信息(PMI)概率评估[8]

互信息是测量两个信息之间相关性的常用方法。互信息的定义为[9]：

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(X \mid Y)$$

$I(X;Y)$ 描述了包含在 X 中的有关于 Y 的信息量，以及包含在 Y 中的有关于 X 的信息量；$H(X)$、$H(Y)$分别表示事件 X、Y 的信息熵。

实际应用中，常使用互信息的方法计算两个具体事件之间的互信息，称为“点互信息”，事件 X、Y 之间的点互信息定义为：

$$I(X;Y) = \log\frac{p(X,Y)}{p(X)p(Y)}$$

$p(X,Y)$表示事件 X 和 Y 同时出现的概率，$p(X)$、$p(Y)$分别表示 X 和 Y 出现的概率。

2)N-gram 算法概率评估[10]

句子 S 可以表示为一个序列 $S = w_1 w_2 \ldots w_i$，通过语言模型，定义句子 S 的概率 $P(S)$为：

$$p(S) = \prod_{i=1}^{n} p(w_i \mid w_1 w_2 \ldots w_{i-1})$$

结合朴素贝叶斯公式，可以对一个序列做出概率评估：

$w_1 w_2 \ldots w_{i-1} = \text{arg}MaxP(S \mid w_1 w_2 \ldots w_{i-1})$，其中 $w_1 w_2 \ldots w_{i-1}$ 即为 φ 最有可能出现的序列。

4 系统测试与结果

4.1 运行环境搭建

为了满足大数据挖掘的需要，使用 6 个服务器组成的 Hadoop 系统集群，运用 map reduce 分布式计算，并使用 hbase 等工具，提供了增量处理、信息实时更新的能力。依据对模块的划分，分成 3 个部分搭建完成：1)网络抓取模块，使用一个爬虫程序作为该模块的核心，对互联网数据进行抓取，完成了对网络站点直接抓取和使用基于特定模板规则的自动定制查询抓取数据的设计。2)搭建了中文自然语言处理模块，使用该模块可以将基本语料处理成为标注完成的 XML 文件。3)将信息预处理模块和信息挖掘系统模块进行合并，使得两者的关系更为紧密，减少信息在两个模块之间相互调用。

在此系统上，对新词发现与词性推断、城市地名发现等两种实际应用进行了实验。

4.2 新词发现与词性推断

从人民日报的官方网站上，下载了 2009～2015 年度的所有电子报纸数据作为语料。随后对其进行自然语言处理，转换成为 XML 格式的文件。使用点互信息的算法对新词之间的关系进行概率评估，通过选取不同的阈值，分析准确率和召回率之间的关系，如图 2 所示。

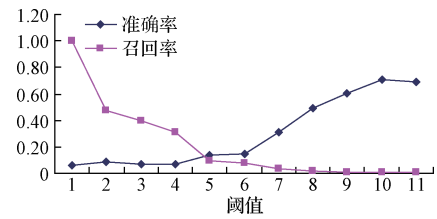

图 2 阈值选取与准确率、召回率之间的关系

分析图 2 可以看出，当阈值为 10 时，准确率达到最高，这充分说明组合词汇基本服从点互信息所描绘的统计规律。选取阈值为 9，对新词进行机器筛选，机器筛选完成后进行人工筛选，把筛选后的词语作为备选新词。接下来，根据新词提取句型模板，使用 Bootstrapping 技术对新词进行词性标注。将备选新词作为诱导种子重新输入系统，系统筛选包含备选新词的语料。

对应不同的备选新词，分析筛选出语料中新词位置的词性，使用分析出来的词性标注备选新词。如果存在多个词性，根据筛选出语料的多少按百分比分配。

词性种类与正确率之间的关系如图 3 所示，如果只取一个词性(含有多个词性时，取百分比最高的词性)，词性标注的正确率为 47.1%。随着词性种类的增加，正确率可以达到 72%。

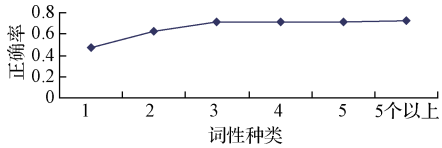

图 3　词性种类与正确率之间的关系

4.3　城市地名发现

使用在新词发现实验中使用的 2009～2015 年度《人民日报》数据作为语料，一方面可以节省处理语料的时间，另一方面，对《人民日报》的语料通过新词发现和词性推断实验后，有效地屏蔽了某些未登录新词对实验结果的影响，减少了误差。实验流程如图 4 所示。

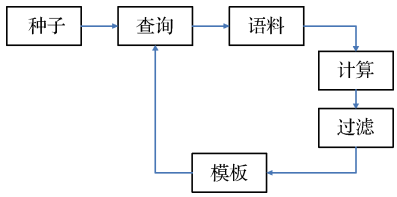

图 4　城市地名发现实验流程

首先通过带有语法限定的诱导种子库，对语料作一个初步的过滤。诱导种子是提取数据的基础，语法限定主要是对诱导种子的补充描述。本实验的主要目的是有效发现城市地名，因此使用“北京”、“天津”、“西安”等作为诱导种子，并使用模板“NeLoc 市”作为语法限定。将诱导种子定义为：{NP 市|NP∈NeLoc 且 NP∈{北京，天津，西安}}

NP 表示过滤到的词，它只能是“北京”、“天津”、“西安”中的一个，且词性必须是 *NeLoc*。

对筛选语料使用 Bootstrapping 技术进行模板提取。在本实验中，结合点互信息和朴素贝叶斯方法进行概率评估，推断出最佳的模板组合，并将其提取出来：

$$f_1, f_2, \ldots, f_n = \text{argMax} P(\varphi \mid f_1, f_2, \ldots, f_n)$$

其中 $f_1, f_2, \ldots, f_n$ 就是要找的候选模板。然后，经过一定的人工筛选，将正确的模板添加到模板库中。模板主要分为两种形式，一种是单一形式的，如：{NP1 的首都是 NP2|NP1∈NeLoc，NP2∈NeLoc}；另一种是队列形式，如{NPList 等地|NPlist 是一个关于地点的队列}。

使用筛选出来的模板对语料进行筛选，提取种子，再将种子进行人工检查后添加到诱导种子库中，继续循环运行。

本实验共循环了 500 次，取阈值为 4.1，筛选结果为 147 条，正确率为 87.28%。

5　结　论

本文利用自然语言处理和分布式计算的方法，设计实现了一种弱监督的知识获取系统，能够有效地对互联网大规模语料库信息进行自动挖掘和获取，并在新词发现与词性推断、城市地名发现两种应用中取得了较高的准确率。然而，系统采用的算法仍然需要进一步完善，计划将该系统进一步用于模板提取、命名实体识别、指代消解等方面，使其成为一个完全开放的、可扩展的、无监督的知识自动获取系统。

参 考 文 献

[1] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Web-scale information extraction in knowitall:(preliminary results)[C]. International Conference on World Wide Web, DBLP, 2004:100-110.

[2] 鲁强，周新. 基于在线检测动态一维下料问题的 GPU 并行蚁群算法[J]. 仪器仪表学报，2015，36(8)：1774-1782.

[3] 颜培清，何炳蔚，雷阿唐，等. 基于深度信息的多目标抓取规划方法研究[J]. 电子测量与仪器学报，2016，30(9)：1342-1350.

[4] 刘兆军. XML 文档数据集聚类问题研究[D]. 长春：吉林大学，2015.

[5] 张鸿鹏，李东新. 动态百分比特征裁剪 AdaBoost 人脸检测算法[J]. 国外电子测量技术，2016，35(9)：37-40.

[6] 郑学伟. 基于语义的信息时序检测技术设计研究[J]. 电子测量技术，2016，39(10)：48-51.

[7] 葛顺，夏学知. 基于聚类的朴素贝叶斯分类无监督学习方法[J]. 舰船科学技术，2016，38(1)：112-116.

[8] DOWNEY D, ETZIONI O, SODERLAND S. Analysis of a probabilistic model of redundancy in unsupervised information extraction[J]. Artificial Intelligence, 2010, 174(11):726-748.

[9] 杜丽萍，李晓戈，周元哲，等. 互信息改进方法在术语抽取中的应用[J]. 计算机应用，2015，35(4)：996-1000.

[10] SIDDHARTHAN A, CHRISTOPHER D, MANNING C D, et al. Foundations of Statistical Natural Language Processing[J]. Natural Language Engineering, 2002, 8(1):91-92.

作 者 简 介

田东，2002 年于中国青年政治学院获得学士学位，2012 年于西安邮电大学获得硕士学位，现为西安邮电大学讲师，主要研究方向为计算机应用技术。
E-mail：tiandong@xupt.edu.cn

张西宁，2004 年于西安邮电学院获得学士学位，2009 年于西北大学获得硕士学位，现为陕西广电网络传媒(集团)股份有限公司工程师，主要研究方向为计算机应用技术。
E-mail：xiningzhang@sina.com