

基于自然语言处理的知识检索算法研究

贾润亮

(太原理工大学 财经学院信息系, 山西 太原 030024)

摘 要: 当前的知识检索存在检索速度慢、效率低的问题, 对此提出一种基于自然语言处理的知识检索方式, 利用半监督算法进行自然语言描述, 针对全部知识检索对象构建无向图, 在该无向图中, 设计一种标签传递方法, 将有标记的数据样本传递到其余没有标记的数据样本中, 完成无向图中所有数据标记的过程. 利用 TextRank 方式, 针对所有标记后的数据选取主题词, 根据检索对象中的主题词实现知识检索的过程. 实验结果表明, 利用此算法进行知识检索, 能够有效提高知识检索的速度, 在短时间内完成知识检索的过程, 提高用户的满意度.

关键词: 自然语言处理; 知识检索; 半监督算法

中图分类号: G252.7

文献标识码: A

文章编号: 1000-7180(2016)10-0130-04

The Knowledge Retrieval Algorithm Based on Natural Language Processing Research

JIA Run-liang

(Department of Information, College of Finance & Economics, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: The current knowledge retrieval is retrieval speed is slow, the problem of low efficiency. Put forward a way of knowledge retrieval based on natural language processing, the use of a semi-supervised algorithms for natural language description, to construct knowledge retrieve objects for all undirected graph, in the undirected graph, transmission method to design a label, there will be passed to the rest of the data sample no marking data samples, complete all the data in the undirected graph marking process. Using TextRank way, in view of the data to select keywords after all the tags, according to the keywords in retrieve objects to realize the process of knowledge retrieval. The experimental results show that the algorithm presented in this paper a knowledge retrieval, can effectively improve the speed of knowledge retrieval, in a short period of time to complete the process of knowledge retrieval, improve user satisfaction.

Key words: natural language processing; knowledge retrieval; a semi-supervised algorithms

1 引言

随着数据挖掘技术的不断发展, 对知识的储备和获取是成为提升工作效率的重要手段^[1]. 在智能数据挖掘技术不断提升的情况下, 对于知识检索有了更加准确、高效、快速的需求^[2,3]. 在网络技术不断提升和网络化程度日益增加的情况下, 网络中的知识检索已经成为人们获取目标知识的重要途

径^[4-6]. 在知识检索的过程中, 用户需要在最短的时间内, 检索到有价值的知识, 并且删除冗余知识^[7]. 与此同时, 用户需要在知识检索的过程中, 获取大量具有足够参考价值的知识, 提高效率, 扩展思路. 自然语言处理是通过计算机的解读, 使知识能够转换为日常语言, 构建用户与计算机的互动关系, 利用上述关系可以完成在大量的网络信息中进行知识检索的目标, 有效提升知识检索的效率和准确度^[8-10]. 根

据网络中大量信息的特性、用户的需求特性,利用自然语言描述目标知识,完成数据检索.因此,自然语言处理方式在知识检索方面得到了广泛的应用,也拥有极大地发展空间和研究价值,成为很多专家和学者研究的焦点.

在知识检索过程中,很多研究者发现,利用自然语言进行知识描述,不仅可以降低用户在知识检索过程中的操作复杂度,还可以提高知识检索的效率和准确度.为了提升自然语言处理在知识检索方面的应用度,同时有效提升知识检索的效率和准确度,本文提出了将半监督算法与 TextRank 相结合的方式,实现基于自然语言处理的知识检索.实验结果表明,利用自然语言处理进行知识检索,可以极大的提升检索的效率与准确度,取得了令人满意的结果.

2 基于自然语言处理的知识检索相关原理

2.1 半监督算法在无向图建立中的应用

半监督算法在特定数据检索中有较大的优势,主要是由于该算法在信息检索过程中,能够更加直观、容易地完成信息匹配,具有较强的可释读性.因此,在诸多不同类型的知识检索中都得到了比较理想的应用结果.

但是,半监督算法一般情况下都具有较高的复杂性,而且是利用直推的方式实现的,因此,无法训练出固定模型,主要是因为知识检索内容的类型发生改变的情况下,该算法的复杂性较强,一般情况下需要 $P(l o^2)$,在知识检索的过程中,大量任务同时存在的情况下,需要逐步提取文档中词汇的特定特征,该方法难以避免数据稀疏的缺陷,因此,需要对半监督算法进行改进,完成标记传播,使基于自然语言的知识检索方式能够大规模应用,方法如下.

在基于自然语言处理的知识检索过程中,利用半监督算法将全部数据进行无向图的构建,在构建的图中,每一个点代表一个数据点,两个数据点之间的连接线用来描述数据之间的相似性.设置 X 描述图中连接线的权值矩阵, x_{jk} 用来描述数据点 j 与数据点 k 之间连接线的权值,假设上述两个数据点之间不存在连接线,则 $x_{jk} = 0$,设置 L 是邻近的数据点.假设数据点 k 在数据点 j 的邻近区域中,则 $x_{jk} = 1$,否则 $x_{jk} = 0$.假设用高斯核对上述关系进行描述,则能够得到下述结论:

$$x_{jk} = \exp\left(-\frac{\|y_j - y_k\|^2}{\zeta^2}\right) \quad (1)$$

在上面阐述的过程中,在流形条件下引入半监

督学习算法.所谓的流形条件,即在数据样本中数据复杂度比较高的情况下,将其存储在数据复杂度比较低的样本中.因此,如果上述流形假设成立,则基于自然语言处理的知识检索方法可以在数据复杂度比较低的样本空间中实现.利用上面阐述的方法进行基于自然语言处理的知识检索,能够避免由于数据复杂度过高带来的知识检索的局限性,充分展示了半监督算法中决策函数的优越性.

因此,邻近区域中的数据样本拥有较强的相似度.基于自然语言处理的知识检索方法计算复杂度也相对比较高,需要针对上述知识检索方式建立与数据点对应的无向图,通过运算获取对应数据的拉普拉斯矩阵,根据获取的结果,得到拉普拉斯特性相关数据,假设上述数据的无向图具有较高的完整度.

2.2 降低知识检索计算复杂度

标记传播方法可以简化无向图,利用该方法,能够有效降低知识检索方式的复杂度,提高知识检索方式的可用空间.在上述过程中,利用流形假设进行数据处理,能够获取数据样本相关的邻近区域图谱或者完整性较高的无向图.在图中,假设部分数据样本是存在标注的,则已经进行标注的数据样本标签能够传递到其余数据样本中.如果数据样本的相似度较高,则上述数据样本的传递难度相对较低.在上述数据样本的标签传递过程中,标签不发生改变,但是没有进行标注的数据样本的标签会进行相应的数据更新.

在根据数据样本构建无向图的过程中,两个数据点的相似度能够进行自定义,可设置利用高斯核进行相似度定义,针对数据样本之间标签传递的过程,能够建立标签传递概率矩阵,设置对应的矩阵大小是 $(M+V) \times (M+V)$,则能够得到下述结果:

$$U_{jk} = Q(j \rightarrow k) = \frac{x_{jk}}{\sum_{l=1}^{M+V} M+V} \quad (2)$$

运用以上方法对无向图进行具体步骤如下所示.

输入:全部数据样本构成的集合 E ,其中包括有标记的数据构成的集合 M 和没有标记的数据构成的集合 V ,不同类别的数据属性构成的集合 D ;

输出:没有进行标记的数据样本的标签;

算法如下:

(1) 根据自定义的数据样本之间的相似度计算方式,构建数据样本的无向图 $H = \langle W, F \rangle$,获取对应的相似度矩阵 X ;

(2) 通过计算获取数据样本标签传递过程中的传递概率矩阵 U ;

(3) 根据数据样本标签传递的概率, 计算数据标签值之和, 并对对应的数据概率分布进行有效的更新, 其公式如下:

$$G_{jk} = \sum_{y=1}^{M+V} U_{jk} \quad (3)$$

(4) 针对已经确定有标注的数据样本, 需要设置该数据点的概率值等于初始值, 不断重复步骤 (3), 直至达到收敛条件。

假设在上述基于自然语言处理的知识检索的过程中, 时间复杂性取值超过 $P(lp^2)$, 由于标记传播方法属于直推方式的知识检索方式, 需要在每次知识检索的过程中针对算法进行再次测试, 所以, 需要在短时间内提取数据样本中的主题词, 降低知识检索的过程中计算的复杂性。

2.3 基于 TextRank 的主题词搜索方式

在基于自然语言处理的知识检索过程中, 需要利用 TextRank 进行主题词搜索。在主题词搜索过程中, 主要针对文档中存在的频率比较高的词语进行研究, 假设某些词语在文档中存在的频率比较高, 则代表这样的词语能够在比较广泛的领域中得到应用。所以, 上述主题词能够直观地反映文档的核心意思。在基于自然语言处理的知识检索过程中, 可以利用 TextRank 针对主题词进行快速、高效、准确的搜索。这种搜索方式是在自然语言处理领域中的基础排序方式。设置 TextRank 模型能够用来描述数据样本构建的无向无权重图 $H(W, F)$, 在该图中, 所有的数据点构成的集合能够用 W 进行描述, 所有的边构成的集合能够用 F 进行描述, 每一个数据样本都与一个主题词相对应, 每一条边都与同现关系相对应。上述过程能够用下述步骤进行描述:

(1) 数据样本初始化处理。在数据库中, 针对所有的文档进行格式转换, 将 PDF 文档转换为 TXT 文档, 同时, 针对所有的文档进行分词与分词标注, 为主题词搜索进行数据准备。

(2) 数据样本过滤。在数据样本过滤的过程中, 首先搜索词语的性质是名词、动词与形容词的词语, 上述词语性质的词语属于主题词的概率比较大, 因此, 需要存储到对应的数据集合中。根据停用词表, 删除所有的停用词, 实现初始的数据样本过滤。

(3) 统计指定词语出现的频率。在基于自然语言处理的知识检索过程中, 需要针对加权的词语出现的频率进行有效的统计, 获取的结果用数据集存储。

在无向图中, 每一个数据节点, 都存在词语、词语出现的频率、词语的性质与索引四个不同的属性。在上述属性中, 索引用来描述数据样本点的空间位置。

(4) 词语合并。针对没有登录的短语, 进行合并处理。

(5) 同现分析。设置一个数据集合中的所有数据构建一个同现窗口, 在该同现窗口中, 假设两个词语在同一时间内出现, 则代表上述两个词语存在同现关系。不断更新上述窗口的空间位置, 直至该窗口出现在指定文档的末端, 此时, 能够获取 TextRank 模型, 在该模型中, 每一个点能够用来描述数据样本中的一个特征词, 每一条线能够用来描述具有同现关系的词汇。

(6) 利用下述公式计算待选特征词的取值, 该取值主要由两个方面获取, 其中, 一个方面是同现关系特征函数, 另一个方面是加权词特征函数。

$$t(w_j) = b \sum_{k \in M(w)} F_{jk} / U + \chi g_j / (1 + g_j) \quad (4)$$

在上述公式中, $t(w_j)$ 能够用来描述数据样本中特征词 w_j 的特征值, 该数据的取值区间是 $[0, 1]$; β, χ 能够用来描述数据的加权重, $M(w_j)$ 能够用来描述与数据点 w_j 联通的所有数据点构成的集合, F_{jk} 能够用来描述数据点 v_j 与 v_k 的连接线, 该连接线的取值是 1。 U 能够用来描述归一化公式的分母, 在针对指定的文档进行知识检索的过程中, 该取值是设定的固定值。假设数据点的数量能够用 $D(w)$ 进行描述, U 的取值能够用 $D(w) * (D(w) - 1) / 2$ 进行描述, g_j 能够用来描述词语 w_j 出现的频率。根据以往的实验经验, 可以进行词语权值的设定, 设置词频的权值是 1, 在文档的标题部分, 词频的权值通常设置为 2, 在摘要中, 词频的权值通常设置为 1.8, 在每一段的开始或者结束部分, 词频的权值通常设置为 1.5。

3 仿真实验结果与分析

为了证明提出的基于自然语言处理的知识检索技术的综合有效性, 需要在在 Matlab7.0 的环境下搭建实验仿真平台, 分别采用基于 TRS 全文检索的知识检索方法和基于自然语言处理的知识检索方法进行实验。

在知识检索对象是 5 GB 内存的情况下, 针对该小规模数据样本集合, 分别利用本文方法和基于 TRS 全文检索的知识检索方法进行知识检索实验, 将两种不同方法进行知识检索的结果对比, 对比结果如图 1 所示。

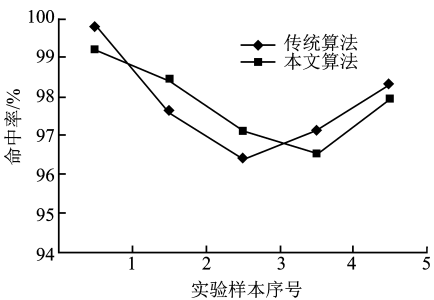


图 1 不同算法知识检索结果

根据上述实验结果能够得知,利用传统算法和本文算法进行知识检索,检索的结果基本趋于一致,代表本文算法与传统算法在知识检索的过程中准确度基本没有差异。

在上述实验的过程中,检索时间能够用表 1 进行描述。

表 1 不同算法在小规模对象中知识检索的对比测试

算法	命中率	词频	检索时间/s
传统算法	36 542	74 516	0.213
本文算法	65 41	11 526	0.211

根据表 1 可以得知,利用本文算法与传统算法进行知识检索,检索时间也非常接近,代表在数据规模比较小的情况下,本文算法与传统算法的检索性能趋于一致。

在实验对象数据规模是 20 GB 的情况下,分别利用传统算法和本文算法进行知识检索实验.实验的结果如图 2 所示。

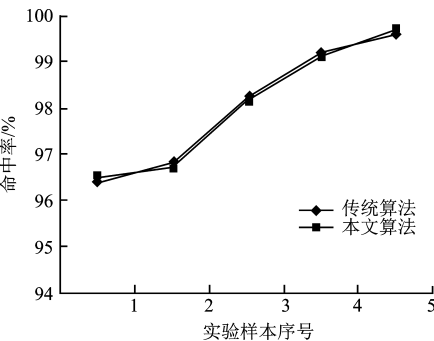


图 2 不同算法知识检索结果

根据上述实验结果能够得知,利用传统算法和本文算法进行知识检索,检索的结果基本趋于一致,代表本文算法与传统算法在知识检索的过程中准确度基本没有差异。

在上述实验的过程中,检索时间能够用表 2 进行描述。

表 2 不同算法在大规模对象中知识检索的对比测试

算法	命中率	词频	检索时间/s
传统算法	545 214	1 023 541	0.813
本文算法	633 218	1 354 126	0.651

根据表 2 可以得知,利用本文算法与传统算法进行知识检索,本文算法的检索时间远远少于传统算法,代表在数据规模比较大的情况下,本文算法在检索效率上拥有较强的优越性。

4 结束语

针对 TRS 全文检索方式在知识检索方面的局限性,本文提出了将半监督算法与 TextRank 相结合的方式,实现基于自然语言处理的知识检索.实验结果表明,利用自然语言处理进行知识检索,可以极大地提升检索的效率与准确度,取得了令人满意的结果。

参考文献:

[1] 白宁超,唐聃,王亚强. 基于主动学习的传统中医症状本体构建方法研究综述[J]. 电子技术与软件工程, 2016,22(7):162-163.

[2] 饶娟. 基于 Ontology 的自然语言智能检索技术研究[J]. 科技信息, 2008,14(33):60-60.

[3] 赵瑞雪. 基于知识的农业信息检索与自然语言控制技术[J]. 图书馆理论与实践, 2005,26(2):59-61.

[4] 龚芳,耿骞,王洋. 自然语言检索中的概念控制[J]. 中国图书馆学报, 2005, 31(4):45-48.

[5] 张伟聪. 2008-2012 年自然语言检索研究[J]. 中华医学图书情报杂志, 2013, 22(7):4-5.

[6] 杨建永,户丽萍. 浅析网络信息过载知识在文献检索课中的融合——以宿迁学院为例[J]. 图书馆界, 2015,35(1):18-21.

[7] 田野. 论各种文献信息检索工具及如何选择正确的检索工具[J]. 赤峰学院学报(自然科学版), 2016, 32(3):123-125.

[8] 赵静,党丽琼. 基于自然语言理解的在线答疑系统设计与实现[J]. 计算机时代, 2015,32(5):10-12.

[9] 李千驹,宋培彦. 基于会话合作原则的知识组织方法及其实证研究[J]. 情报理论与实践, 2015, 38(8):91-94.

[10] 王芳. 自然语言逻辑的产生与人工智能中的知识表示[J]. 福建农林大学学报(哲学社会科学版), 2015, 18(5):98-101.

作者简介:

贾润亮 男,(1973-),硕士,讲师.研究方向为人工智能.
E-mail:381095759@qq.com.