基于语义的文本资源分类①

富 宇, 石金叶

(东北石油大学 计算机与信息技术学院, 大庆 163318)

摘 要: 进入 21 世纪以来, 知识数据大量存储在文档中, 但各类文档的粒度和结构不便于知识的加工、整合和管理. 如何从这些无序的、非结构化的数据(知识)源中提取语义, 首要任务是将蕴藏在数据、信息中的知识抽取出来, 建立文本资源的语义网, 采用 RDF 来表示语义数据, 其次采用 TFIDF 算法计算得出文本特征词的可信度, 最后将文本信息录入到数据库中, 实现文本类资源的自动分类, 最终目的是实现文本资源知识的共享.

关键词: 语义网; RDF; TFIDF; 分类

Text Classification Based on Semantic

FU Yu, SHI Jin-Ye

(Institute of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Since the 21st century, much of the knowledge is stored in the document, but the size and structure of all kinds of documents do not facilitate knowledge processing, integration and management. How to extract semantics from the disordered, unstructured data (knowledge) is a task, the primary task is to extract knowledge which is contained in the data and information, to construct the semantic web which is based text resource, using RDF to represent indicate the semantic data, then using TFIDF algorithm to calculates the credibility of the key of the text, the last to enter text information in the database and to realize automatic text classification, ultimate goal is to realize knowledge sharing.

Key words: semantic web; RDF; TFIDF; classification

随着信息技术和网络技术的不断发展, 学习资源出现了规范不一、组织异构、缺乏语义关联等问题, 制约网络教学资源的管理和知识共享. 因此, 新一代互联网(语义网络)被提出并在不断发展中, 它旨在对信息从语义层次上进行组织, 以提高对信息的利用程度.

1 语义网相关理论

为了对数据进行有效地组织和利用, W3C 主席 Tim Berners-Lee 在《编织万维网》一书首次提出了语义网[1], 语义网的基本思想是提供基于机器可处理的数据语义, 并应用这些元数据的启发式进行自动化的信息访问[2]. 其最终目标是将人类知识编织成一个巨大的网络, 并以机器处理的方式来实现它[3]. 所以, 语义网技术并非互联网的替代, 而是一种有益补充[4]. 因此语义技术必定会成为未来现代信息系统的数据基础, 为物联网、云计算等新兴技术的发展提供重要的数据支持[5].

语义网是通过概念及其语义关系来表示知识的一种网络图,它是一个带标注的有向图. 在语义网中,各个节点代表概念、事物等,节点通过描述各节点间相互关系的弧连接. 可以用不同的方法定义弧,这取决于被表达的知识种类[6]. 发展语义 Web 的两个关键技术已经形成: 可扩展标记语言(XML)和资源描述框架(RDF). 这里的 XML 不涉及网页的具体内容, XML 允许每个用户创建自己的标记并任意增加结构, 且无须说明其结构的含义, 而 RDF 则用以表达网页的内容[7].

① 收稿时间:2015-11-29;收到修改稿时间:2016-01-04 [doi:10.15888/j.cnki.csa.005263]

2 RDF概述

W3C 于 1999 年公布的资源描述框架 RDF[8], RDF 提供了一种用于表达语义信息、并使其能在应用程序间交换而不丧失语义的通用框架,是语义网表示语义信息的基础[9]. RDF 用主语、谓词、宾语的三元组形式来描述 web 上的资源. 其中,主语一般用于表示 Web 上的信息实体(或者概念),谓词描述实体所具有的相关属性,宾语为对应的属性值, 这样的表述方式使得 RDF 可以用来表示 Web 上的任何被标识的信息[10],并且使得它可以在应用程序之间交换而不丧失语义信息. 因此, RDF 成为语义数据描述的标准[11].

采用 RDF 三元组形式构建文本特征词的知识结构体系, 分析、归纳文本中特征词以及特征词间的关联关系, 可以有效的表示文本类资源中特征词之间的关系, 同时也可以使各个文档之间的文档粒度降到最低, 有利于从非结构化的文档中提取语义, 降低各文档之间的粒度, 同时提高查询效率和查询相关资源的准确性.

3 文本资源分类研究

文本是非结构化的数据, 要想从大量的文本中挖掘有用的信息就必须首先将文本转化为可处理的结构化形式, 然后找出对文本特征类别最具代表性的文本特征, 再通过统计方法计算术语的可信度. 本节会选取电脑相关的文档进行实验, 本实验分四部分: 基于语言规则获取候选文本词汇、采用 TFIDF 计算候选文本词汇的可信度、将候选文本词汇与语义图进行语义匹配, 最后获取查询页面.

3.1 实验及结果

3.1.1 利用 RDF 技术进行语义数据的描述

本文选取了 CCL 语料库中关于电脑的相关文档进行实验. 构建相关语义图如图 1 所示.

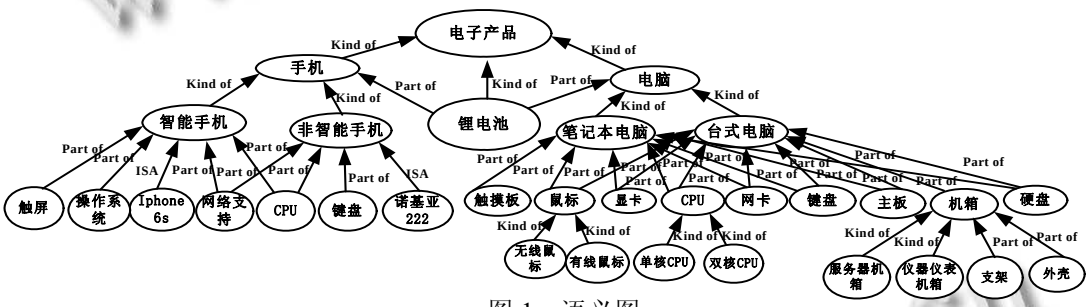

图 1 语义图

3.1.2 基于语言规则获取候选文本词汇

利用 ICTCLAS 切词工具对选取的文档进行分词, 分词结果如图 2 所示.

1 作为/p 电子/n 产品/n 的/u 一部分/m,
2 电脑/n 分为/v 台式/n 电脑/n、笔记
3 本/n 电脑/n ， 电脑/n 主要/n 包括/v：
4 CPU/n、鼠标/n 、硬盘/n、键盘/n、
5 屏幕/n、显卡/n 、机箱/n、内存条/n、
6 主板/n 等/u, 其中/d CPU/n 分/v 单核/n
7 和/p 双核/n 等/u, 电脑/n 给/v 人们/n
8 带来/v 了/y 便利/n.

图 2 ICTCLAS 工具分词结果

由于分词工具的分词结果主要是基本词汇, 但是领域中组合词汇所占的比重也比较大. 如果直接对基本词汇进行概率统计以获取候选术语可能会丢失一些重要信息, 针对这种情况, 本文根据 ICTCLAS 的词性标注定义了抽取组合短语的规则: . 即规定短语可以由多个形容词、名词或者动名词组成, 但是必须以名词结尾, 基于这一原则实验证明该规则
能够有效地获取名词性短语. 基于规则匹配的方法获取组合词汇, 实验结果如图 3 所示.

1 电子产品/n 电脑/n 台式电脑/n 笔记本电脑/n 电
2 脑/n CPU/n 鼠标/n 硬盘/n 键盘/n 屏幕/n 显卡/n
3 机箱/n 内存条/n CPU/n 单核/n 双核/n 电脑/n

图 3 获取组合词汇结果

在基于规则重新组合之后, 如“电子产品”, “笔记本电脑”, “台式电脑”等这样的组合词汇(文本特征词)均被抽取出来, 对文本资源的分类起关键性作用.

3.1.3 文本特征词的领域可信度

TFIDF 是一种加权技术, 它通过统计的方法来计算和表达某个关键词在文本中的重要程度, 该方法以特征独立的假设为基础, 能够简化特征词提取, 降低计算时间. TF(Term Frequency)表示词语在文本中出现的频率, 词语对文本表示贡献大小由 TF 决定, IDF(Inverse Document Frequency)表示倒排文本频率, 体现词语在文本集中的分布情况. 算法公式如下:

$$TFIDF=\sum_{i=1}^{n} tf_{ij}*\log((1+n))/df_i \quad (1)$$

公式中 tf_{ij} 表示在文件 f_j 中词汇 t_i 的数量, df_i 是包含词汇 t_i 的文件的数量. 由此我们看出, 词汇在文件中出现的次数越高, 其中 TF/IDF 值就会越大, 说明该词汇重要性越大, 相反, 如果含有它的文件的数目越多的话, 其 TF/IDF 的值就会变小, 也就是说明该词汇在各文件中是泛泛呈现, 其领域重要性将会变小. 还有基于词汇信息熵的统计方法, 算法公式如下:

$$ENTR(t)=-\sum_{i=1}^{n} p(c_i|t)\log_2 p(c_i|t) \quad (2)$$

词汇的信息熵表示他在文件中的散布情况, 如果熵值越大则它将会在各个文件中以匀称的状态出现. 相反, 如果熵值越小则表明该词汇的呈现状态越收聚, 可能只是出现在特别的文件中. 另外, C-value 算法主要用于抽取那些被其他术语包含的候选术语. 算法公式如下:

$$C-value=a\log|a|*\left(f(a)-\sum_{b\in T_a} f(b)\Big/ n_T\right) \quad (3)$$

其中 $f(*)$ 实短语或术语出现的频率, T_a 是包含术语 a 的短语集合, n_T 是 T_a 中短语的数量, 同时该方法还考虑了领域术语的长度 $|a|$, 因为较长的术语包含的领域信息越全面, 相对来说也就越重要. 依据上述算法计算由图 3 中得出的组合词汇的可信度, 如表 1 所示.

表 1　特征词的可信度

名词词汇	可信度
电脑	0.262515
台式电脑	0.234142
笔记本电脑	0.193680
电子产品	0.103245
硬盘	0.092283
CPU	0.060989
机箱	0.042567
主板	0.014217
键盘	0.012985
鼠标	0.006325
显卡	0.003563
双核	0.003130
单核	0.001623
屏幕	0.001611
内存条	0.001421

3.1.4 语义匹配及查询结果

语义匹配分为: 第一步, 特征词与语义网中某节点的匹配; 第二步, 特征词与语义网中子网的匹配. 即通过切词工具, 将文本资源进行切词获取文本特征词术语集, 通过 TFIDF 算法获取文本特征词的可信度, 但是仅根据 TFIDF 计算的得出的文本特征词的可信度, 可能忽略掉那些出现频率较低但对文本资源分类为关键作用的词汇, 所以为了提高查询的准确性, 可根据对抽取结果做上文的权重过滤, 即可以根据两个指标对特征词的抽取效果进行衡量: 准确率: $\frac{a}{a+b}*100\%$, 召回率: $\frac{a}{a+c}*100\%$, 其中 a 代表抽取词汇的总数目, b 代表抽取结果中不应该抽取的词汇数目, c 代表应该抽取的词汇但没有被抽取的词汇数目, 然后根据再进行语义匹配, 最终得出匹配节点的可信度, 如图 4 所示.

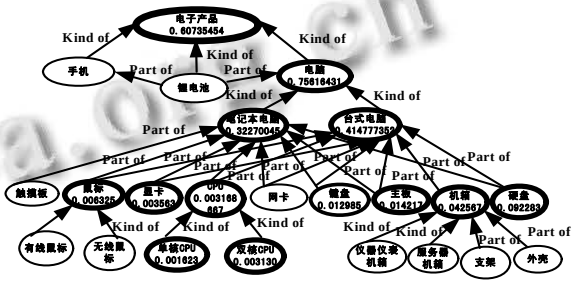

图 4　语义匹配图

图 4 中颜色较深的节点是与之匹配的节点, 图中匹配节点下面的数字为匹配节点的可信度, 根据公式(4)计算得出,

$$f(i)=\sum_{i=1}^{n} T_i*\frac{n}{m+1}+P \quad (4)$$

计算相关节点的可信度, 公式中 P 为根节点的可信度, $\sum_{i=1}^{n} T_i*\frac{n}{n+1}$ 为下层支撑节点的领域可信度的对上层节点的贡献率, 相关匹配节点的可信度由图 4 得出,

其中 $\sum_{i=1}^{n} T_i$ 为支撑节点的可信度的总和, n 为支撑节点的总数, 最终得出节点的可信度, 例如, 台式电脑的可信度由原来的 0.234142 根据公式(4)变为 0.414777352,

f(i)=(0.006325+0.003563+0.060989+0.012985+0.014217+0.092283+0.042567)*7/9, 其他匹配节点的可信度如表 2 所示.

表 2 可信度排序结果

节点名称	可信度
电脑	0.75616431
电子产品	0.60735454
台式电脑	0.414777352
笔记本电脑	0.32270045

根据表 2 中节点的可信度排序结果可知, 该文本在电脑, 电子产品等四方面都有涉及, 即可将文本的相关信息, 如文本名称, 可信度, 分类类别等内容录入到数据库的相关表中, 以供用户查询. 即当用户输入在查询页面输入“电脑”(数据库中存储的类别名称)后会返回用户一个查询页面, 包括: 文件名称, 可信度, 作者, 时间等信息, 如图 5 所示.

图 5 页面查询结果

根据查询页面结果可浏览、下载与电脑相关的文本资源, 实现文本资源知识的共享, 同时可根据相关查询条件在数据库相关表中进行 SQL 查询, 得出与之相应的查询结果, 即与查询关键词匹配的文本名称, 作者, 上传时间等字段名称的信息.

4 结语

本文通过采用 RDF 来描述文本资源, 有利于后期文本特征词之间的关联和编织. 通过切词, 特征词提取的算法, 可以将蕴藏在数据、信息中的特征词抽取出来, 依据其可信度, 进行语义匹配, 获得与之匹配的相关节点信息, 实现文本资源的分类, 最终在学习平台上实现文本资源的查询, 为学习者提供准确、及时的文本资源, 实现了文本类知识资源的共享.

参考文献

1 Berners-Lee T. Weaving the Web, Harper, San Francis-co, CA, 1999.

2 朱礼军,陶兰,黄赤.语义万维网的概念、方法及应用.计算机工程与应用,2004,(3):25–26.

3 Berners-Lee T, Hendler J. Publishing on the semantic web, Nature 410, 26 April 2001: 1023–1024.

4 Gu ZF, Xu B. Service data correlation modeling and its application in data-driven service composition. IEEE Trans. on Services Computing, 2010, 3(4): 279–291.

5 陈华钧,谢国彤,潘越.语义万维网的应用.中国计算机学会通讯,2010,6(8):30–37.

6 Wang Z, Djuric N. Trading representability for scalability: adaptive multi-hyperplane machine for nonlinear classificatiotin. Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego. 2011,13(8). 22–32.

7 白同强,刘磊.语义 Web 的研究与展望.吉林大学学报, 2004, 22(2):154–158.

8 Fenseld. The semantic web and its language. IEEE Computer Society, 2000, November December,(6):67–73.

9 易雅鑫,宋自林,尹康银.RDF 数据存储模式研究及实现.情报科学,2007,25(8):1219.

10 RDF concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/.

11 杜方,陈跃国,杜小勇.RDF 数据查询处理技术综述.软件学报,2013,24(6):1222–1242.