

供需信息跨语言检索算法研究

姚寒冰,王丽清,徐永跃

(云南大学 信息学院 云南省高校数字媒体技术重点实验室,云南 昆明 650223)

摘要:经济全球化促进了互联网电子商务的快速发展,跨境电商因其巨大的发展潜力成为新的贸易增长点。由于贸易的基础与前提是供需双方信息的高效共享和沟通,而跨境电商因涉及不同语言之间的互译,使得信息交流的及时性、准确性不足,导致丧失贸易时机,甚至导致贸易失败。为此,提出了一种基于自然语言的跨语言协同机器翻译的信息检索算法。该算法可使供给方可根据所提供商品服务的特点进行灵活的扩展描述,并为需求方提供自然语言描述方法,需求方可使用不同的语言进行输入,完成跨语言的检索。为验证协同机器翻译的自然语言实现供需信息的检索和自动匹配能力,进行了相关验证实验测试。实验测试结果表明,所提出的算法可满足供给方对自身商品或服务进行特有属性扩展描述的需求,同时具有多语种拓展潜力,有助于消除供需双方的语言障碍。

关键词:跨语种;供需;自然语言;检索

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2017)08-0152-04

doi:10.3969/j.issn.1673-629X.2017.08.032

Research on Automatic Retrieving Algorithm of Cross-language Supply and Demand Information

YAO Han-bing, WANG Li-qing, XU Yong-yue

(Key Lab of Digital Media Technology of Universities in Yunnan Province, School of Information Science and Engineering of Yunnan University, Kunming 650223, China)

Abstract: Economic globalization has promoted the rapid development of Internet e-commerce. And the cross-border e-commerce suppliers have become a new growth point of trade due to its huge potential of development. Since basis and prerequisite of trade are efficient sharing and communication of information between supplier and demander, cross-border e-commerce involves translation between different languages, which results in lack of the real-time and accuracy in information exchanges as well as miss of trade opportunities or even failure in trade. To solve this problem a cross-language information retrieval algorithm based on natural language and collaborated with machine translation is proposed, which enable supplier to describe the characteristics of the goods or services flexibly and provide demander for different natural language to describe its own demand for completion of retrieval of cross-language. In order to verify ability of retrieval and automatic matching of supplier and demander information collaborated with machine translation, the verification experiment has been conducted and its results show that the proposed algorithm has met demands of the supplier to depict unique attributes of their goods or services and has owned potentials for multilingual development, which can help to eliminate the language barrier between the supplier and demander.

Key words: cross-language; supply and demand; natural language; retrieval

0 引言

近年来,对于供需自动检索的方法和跨语言信息检索已有很多研究。最常见的检索方式是对供给方的商品或服务,进行树状目录分级,由需求方逐级进行人工选择。这种方式操作方便,但是存在供给方的商品

或服务只能套入固定模式的树状目录分级结构、难以扩展某些独特商品或服务的特色,以及需求方不一定对于该树状目录分级结构很清楚等弊端。另外,还有基于多目标离散差分进化算法的交易检索方法^[1]、基于 B2B 电子交易环境的供需匹配概念框架^[2]、基于图

收稿日期:2016-08-18

修回日期:2016-11-23

网络出版时间:2017-07-05

基金项目:云南省科技创新强省资助项目(2014AB021)

作者简介:姚寒冰(1978-),男,工程师,硕士,研究方向为信息系统;王丽清,通信作者,副研究员,硕士生导师,研究方向为信息系统与检索、电子商务等。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20170705.1650.038.html>

论的商品自动匹配系统^[3]、基于商品本体结构语义相似度匹配算法^[4]、基于电子中介下商品交易为背景的方法^[5]等。但是以上方法,都没有解决跨语种供需信息自动匹配检索的问题,导致不同语种环境下的应用存在局限。

自然语言检索方面的研究包括面向自然语言检索的标引技术、自然语言提问分析与处理、自然语言检索的匹配过程及概念控制^[6]、基于有限状态方法模型的自然语言处理^[7]、基于语义的自然语言检索方法^[8]等;多语种信息组织与检索方面的研究包括多语言本体构建与协调、基于关联数据的多语言语义网建设、跨语种语言资源和知识组织系统互操作、多语言文本分类与聚类、交互式多语言信息检索^[9]、基于聚类的个性化跨语言信息检索方法^[10]、基于知识源、双语词典和机器翻译的跨语言检索^[11]、在语境单元框架上的匹配和生成机制实现跨语言检索^[12]、基于可对比语料库训练的跨语言信息检索模型^[13]等,还有基于语义网的多语种自然语言查询方法^[14]等。

为此,提出了一种基于语义分析的信息检索算法,即由需求方输入一段自然语言描述需求信息,并与供给信息进行比对。该算法在对需求方的自然语言提问进行语义分析处理、对多语种结构组织的信息库进行匹配检索、对中间库和同义词库进行共同检索的基础上,按照权重算法进行共有特征(包括颜色、重量、价格等商品特有属性)的检索比对,借助人工/机器翻译机制,建立多语种的供求商品或服务信息库,该库作为跨语言检索的中间库,同时构建同义词库提供比对。

1 设计与实现

1.1 供求商品或服务信息库设计

供求商品或服务信息库用于提供供方商品或服务的有关信息,由以下几部分组成:

- (1) 供给信息描述。
每一大类商品或服务,设置相对固定的一系列属

性,即固定属性。对于每一单独的商品或服务,另可各自扩展一系列不确定总数的属性,称为自定义属性,由多个可准确描述商品或服务的独具特点的词汇构成。属性分为文本类型和数值类型,数值类型的属性,还需提供单位名称,并可有上下限。文本类型的属性值、单位名称,由人工/机器翻译机制取得多语言结果并完成存储。

- (2) 自定义属性。
自定义属性,如果由每一单独商品或服务各自分散存储,结果将极大地增加数据库冗余,进而降低检索算法的效率,恶化用户体验。因此,在实现中进行集中存储,即多个相同的自定义属性值,只存储一条。
自定义属性与商品或服务之间构成多对多的关联关系。1 种商品或服务可具有 1 条或多条自定义属性,1 条自定义属性值可归属于 1 种或多种不同商品或服务。

- (3) 权重。
对于固定属性和自定义属性,都具有不同的权重值。对于固定属性,预先设置相对固定的权重值。对于自定义属性,预设权重值随系统平台提供的商品服务的不断变化而动态变化,表示该自定义属性的稀有程度,越稀有的自定义属性,权重值越高。

- 权重值的生成是指该商品服务类别中的自定义属性总数与该自定义属性所属商品或服务数量之间的比值。
- (4) 同义词表。
同义词表用于完成含义相似、相近词的检索,获取一致的结果。

在一个语种中,一组含义相同或相近的词汇可构成一组同义词。基于各语种的同义词典,以及相关商品或服务领域的专业知识,构建同义词表。

1.2 库生成和更新

供求商品或服务信息库的生成和更新,由不同角色协同完成,如图 1 所示。

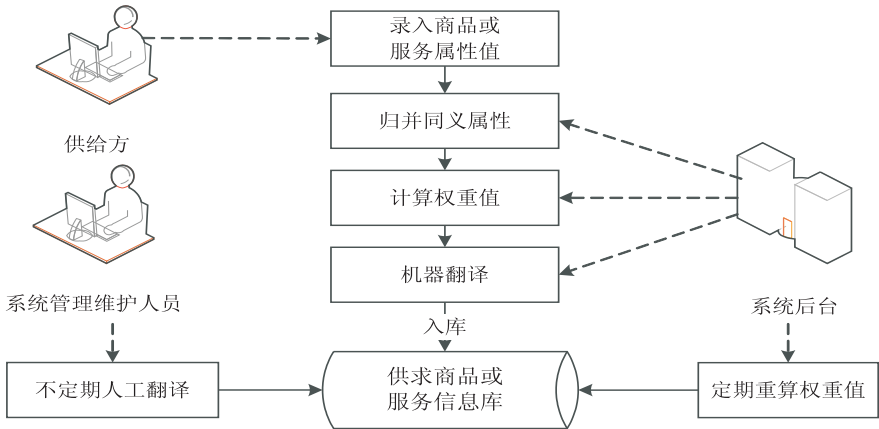


图 1 供求商品或服务信息库生成和更新流程

如图 1 所示,新入库的商品或服务,由供给方人工录入其固定属性和自定义属性;系统后台自动根据当前语种检索同义词典,归并同义属性,并计算权重值;系统后台自动由机器翻译得到其他语种对应属性值;系统后台自动对所有已录入的自定义属性定期扫描以重新设置权重值;系统管理维护人员不定期人工检查机器翻译得到的属性值,并进行人工翻译校正。

供求商品或服务信息库的构建,以中文为主。信息库构建完成后,便可根据一种语种的检索匹配,迅速找到所有已有语种的对应信息库内容,为供给方-需求方语言不通情况下的供需匹配提供一定的便利。通过数据库表扩展字段,即可支持新语种的加入。实现了跨语种和可扩展的特性。并通过关键词和同义词的关联关系,提高检索匹配信息库内容的准确性和兼容性。

1.3 跨语言信息检索算法

当需求方提出需求信息时,由跨语言信息检索算法对供求商品或服务信息库进行检索,实现供需匹配。算法实现的主要思路是:需求方输入的需求信息,与供求信息库中的商品或服务的属性值进行比对,命中的属性权重值总和,超过一定阈值时,即为匹配成功。按权重总和由高到低进行排列,表示匹配程度的吻合度。

在计算过程中,根据属性不同的值类型,有不同的命中定义。

(1)对于文本类型的属性,当需求信息包含该属性值,或者此属性值的同义词时,即为命中。

(2)对于数值类型的属性,根据不同语言的不同表达方式构建不同的正则表达式,形成正则表达式库,并附加该属性的单位,对需求信息进行语义分析,取得数值范围。例如:“300 到 500 元”、“400 元左右”,正则表达式分别为 $\wedge-?[1-9]\d{*}-?[1-9]\d{*}\text{元}\$, \wedge-?[1-9]\d{*}\text{元左右}\$$ 。

(3)对于数值类型的属性,对需求信息中不同形式的单位描述,设置单位换算规则,如需求信息描述与供应信息所使用的单位不符时,可进行换算。

(4)当取得具有上下限的数值范围时,属性值处于该范围之内,即为命中。

(5)当只取得一个数值时,浮动上下 30% 并取整,作为上下限。

这样,需求信息与供给信息匹配程度的吻合度,与命中属性总数、命中属性的稀有性成正比,并能适应自然语言中的不同表达。

具体示例:如权重值总和的阈值为 1 000,供给商品或服务信息库中有某种苹果具有以下固定属性:产地:市(权重值 600);品种:红富士(权重值 350);果径:80~85 mm(权重值 200);是否有机食品:否(权重

值 50);规格:4 000 g(权重值 50);数量:15 个(权重值 50);价格:65 元(权重值 400)。具有以下自定义属性:套袋防虫(权重值 700)。

当有需求方提交需求信息:“A 市产的有套袋防虫的红富士苹果,每公斤 15 元左右。”A 市、红富士、套袋防虫 3 个属性由于被文本包含而命中,并由语义分析获得价格需求:15 元、单位:公斤,供求信息中的规格为 4 000 g,根据单位换位规则得到需求方的价格需求为 60 元,在商品价格属性浮动范围内,也命中。因此,共命中产地、品种、价格、套袋防虫 4 个属性,权重值总和为 2 050,超过阈值,供给信息和需求信息检索命中,获得了匹配。

当有需求方提交英文信息:“Red fuji apple in A-City, 15 Yuan per kg.”也可通过英文关键词 Red apple、A-City、Yuan、kg 命中有关属性,从而获得检索匹配,这样就可实现跨语种检索,在一定程度上克服供给方、需求方之间的语言障碍。

2 结果分析

算法实现的实际效果,主要依赖于供求商品或服务信息库的建设质量,由以下因素构成:商品或服务信息的总量、商品或服务的属性描述的准确性、同义词库的准确性、商品或服务信息的翻译质量。其中,总量、翻译质量可以用量化指标表示,翻译质量以机器翻译所占的比例代表,比例越高,翻译质量越低。

在应用系统中,基于以上指标,对算法效果进行了测试。测试样例,分别基于中、英、泰三个语种,使用 100 条自然语言描述的需求信息对供给信息库进行匹配检索。

测试前,分别抽取 20 条需求信息样本,人工在供给信息库中逐条分析是否含可匹配的供给信息,得到期待匹配比例,用于与应用系统实际得到的匹配结果的比例进行对比。另外,测试表明检索平均耗时不大于 10 ms,可以满足用户体验要求。

检索得到的匹配结果对比如表 1 所示。

表 1 应用系统匹配效果测试结果						%
语种	商品信息(总量:1 068 种)			服务信息(总量:246 种)		
	信息库机 翻比例	期待匹 配比例	实际匹 配比例	信息库机 翻比例	期待匹 配比例	实际匹 配比例
中文	2	85	63	2	60	38
英文	25	85	53	25	60	32
泰文	60	85	39	60	60	22

由测试结果可以得出,信息库的建设质量对检索效率有较大影响,通过信息库的不断完善,可以满足并改善用户体验。具体方法有增加商品或服务信息的总量,加快人工翻译的进度,增强人工翻译的质量,增加

同义词库的容量和准确性,通过系统界面信息或系统后台人员与供给方的互动沟通等方式引导供给方增强商品或服务信息的准确性。

3 结束语

为解决自然语言实现供需信息的检索和自动匹配,满足供给方对自身商品或服务的特有属性扩展进行描述的需求,提出了一种基于可扩展多语种供求商品或服务信息库和协同机器翻译自然语言的供需信息跨语言信息检索算法。测试结果表明,该算法一定程度上满足了供需信息检索与自动匹配的需求,弥补了传统供需检索匹配方式在自然语言和特性描述支持上的不足,可方便地进行多语种的扩展,使得供需双方的语言障碍在一定程度上得以克服,并获得了较好的用户体验效果。

参考文献:

[1] 蒋忠中,樊治平,汪定伟,等. 具模糊信息的多数量多属性电子交易匹配问题[J]. 管理科学学报,2014,17(5):52-65.

[2] Alpar F Z. Matchmaking framework for B2B e-marketplaces[J]. Informatica Economica Journal,2010,14(4):164-170.

[3] 陈向,刘义,柴跃廷. 基于图论的电子易货商品自动匹配系统[J]. 计算机工程,2009,35(17):283-284.

[4] 陈冬林,聂规划,刘平峰. 基于本体的 B2B 电子商务 MAS 模型及商品匹配算法[J]. 计算机工程与应用,2007,43(10):199-201.

(上接第 151 页)

量的估算方法。以服务器的 CPU 使用率与系统响应时间作为衡量标准,对线程池容量估算方法进行了有效性测试。结果表明,采用该方法的服务器在各种负载条件下均能满足有效的 CPU 利用率与稳定的系统响应时间,既保证了整体稳定性,又能提供稳定的吞吐量。

参考文献:

[1] 张垠波. 线程池技术在并发服务器中的应用[J]. 计算机与数字工程,2012,40(7):153-156.

[2] Ling Yibei, Mullen T, Lin Xiaola. Analysis of optimal thread pool size[J]. ACM SIGOPS Operating Systems Review,2000,34(2):43-45.

[3] 李昊,刘志镜. 线程池技术的研究[J]. 现代电子技术,2004,27(3):77-80.

[4] 詹新林,王公亭,徐晓钟. 基于线程池数据分析系统的设计与实现[J]. 微计算机信息,2008,24(33):266-268.

[5] 张尧学,宋虹,张高. 计算机操作系统教程[M]. 第 4 版. 北京:清华大学出版社,2013.

[5] 梁海明,姜艳萍. 一种考虑中介交易态度的买卖双边匹配决策方法[J]. 运筹与管理,2013,22(5):128-133.

[6] 耿骞,赖茂生. 自然语言检索的实现及其关键问题[J]. 情报科学,2007,25(5):733-741.

[7] Anssi Y J, Andras K, Jacques S. Finite-state methods and models in natural language processing[J]. Natural Language Engineering,2011,17(2):141-144.

[8] 谢文亮,王石榴. 基于语义 Web 的科技期刊网络信息检索及其应用[J]. 科技管理研究,2015,35(2):196-200.

[9] 司莉,庄晓喆,贾欢. 近 10 年来国外多语言信息组织与检索研究进展与启示[J]. 中国图书馆学报,2015,41(4):112-126.

[10] 庞观松,张黎莎,蒋盛益. 个性化跨语言学术搜索技术研究[J]. 情报学报,2011,30(8):870-874.

[11] 张玥杰,郭依昆,连理,等. 基于英汉机译实现跨语言信息检索[J]. 小型微型计算机系统,2004,25(7):1135-1140.

[12] 吴晨,缪建明,张全. 跨语种信息检索中的文本比较及结果生成算法[J]. 计算机工程与应用,2005,41(29):11-15.

[13] Vulic I, Smet W, Moens M F. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora[J]. Information Retrieval, 2013,16(3):331-368.

[14] Al-Nazer A, Albukhitan S, Helmy T. Cross-domain semantic web model for understanding multilingual natural language queries; english/arabic health/food domain use case[J]. Procedia Computer Science,2016,83:607-614.

[6] 封玮,周世平. Java 中的线程池及实现[J]. 计算机系统应用,2004(8):16-18.

[7] 王俊峰. 一种实时集群系统负载均衡通用模型的研究及应用[D]. 长沙:湖南大学,2008.

[8] 欧昌华,李炳法. 线程池在网络服务器程序中的应用[J]. 信息技术,2002(5):11-14.

[9] 宋立昊. 基于线程池的 WEB 服务器实现和监测[D]. 长春:吉林大学,2011.

[10] 王华,马亮,顾明. 线程池技术研究与实现[J]. 计算机应用研究,2005,22(11):141-142.

[11] Pyarali I, Spivak M, Cytron R, et al. Evaluating and optimizing thread pool strategies for real-time CORBA[J]. ACM SIGPLAN Notices,2001,36(8):214-222.

[13] 刘雪梅. 服务器端软件性能分析和诊断[M]. 北京:北京邮电大学出版社,2011.

[14] Belkin R. Mechanism for obtaining a thread from, and returning a thread to, a thread pool without attaching and detaching; U. S. , 6 766 349[P]. 2004-07-20.

[15] 许俊奎,徐凤刚,潘清. Web 服务器性能测试工具的设计与实现[J]. 计算机测量与控制,2005,13(11):1204-1206.