DOI:10.16208/j.issn1000-7024.2014.12.050

基于领域本体的网页主题相关度计算

侯超昆，李石君

（武汉大学 计算机学院，湖北 武汉 430072）

摘 要：为提高网页内容与特定主题之间相关度计算的准确度，提出一种基于领域本体的网页主题相关度计算模型 OBWTCCM（ontology based webpage-topic correlation calculation model）。使用领域本体刻画主题，通过计算本体概念间的语义关系提取主题概念并构造主题语义矩阵，将特征词的统计信息与该矩阵相结合计算网页与主题之间的相关度。该模型改进了向量空间模型在相关度计算时对特征词语义层次分析的不足。实际项目应用结果表明，使用该方法计算得到的网页主题相关度与领域专家的判断总体相符，具有较理想的准确度。

关键词：本体；主题；语义；主题概念；相关度计算

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2014）12-4344-06

Calculation of webpage-topic correlation based on domain ontology

HOU Chao-kun, LI Shi-jun

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract: To improve the accuracy of the correlation calculation between the webpage and a specific topic, a webpage-topic correlation calculation model (OBWTCCM) based on the domain ontology was proposed. The topic was described using the domain ontology, and a topic semantic matrix was built after extracting the topic concepts by computing the semantic relation between the concepts in the ontology. Then the correlation between the webpage and the topic was calculated by combing the matrix and the statistics information of feature words. This model improves the vector space model by adding the consideration in the semantic level. The application of the method in the real project indicates that the result overall fits the judgments of the domain experts and has a satisfied accuracy rate in the correlation calculation.

Key words: ontology; topic; semantic; topic concept; correlation calculation

0 引 言

相关度计算是信息检索、文档分类和聚类、文档管理、推荐系统等诸多领域的关键技术之一。计算网页内容与特定主题间的相关度作为主题爬取的核心步骤，能够帮助专业人员在庞大的互联网信息中找到与特定领域相关的内容，在网络信息过滤与情报发现中具有重要意义。目前计算相关度最常采用的是向量空间模型（vector space model，VSM），一些研究者也提出了一些基于 VSM 的相关度计算模型。然而，这些方法归根到底都是基于词频的统计信息，缺乏对特征词在语义层次上的分析，容易导致相关度计算结果与实际情况间产生较大的偏差。

为了提高网页与特定主题相关度计算的有效性与准确度，本文综合考虑现有模型的优缺点，将领域本体引入计算中，提出了一种基于领域本体的网页主题相关度计算方法。该方法使用领域本体刻画主题，通过将构建的主题语义矩阵与特征词统计信息相结合来进行相关度计算，弥补了传统 VSM 模型过度依赖词频信息而在语义层次分析上存在不足的问题。最后通过实际应用检验该方法的性能。

1 相关研究

1.1 相关度计算模型

在相关度计算中应用最普遍的是向量空间模型 VSM，它将文档 d_j 表示为一个空间向量（$(t_1, w_{j1}), (t_1, w_{j1}), \cdots, (t_1, w_{j1})$），其中 t_i 是文档 d_j 中的特征词，w_{ij} 表示采用统计方法计算得到的 t_i 在 d_j 中的权重。这样，可以通过相似函数计算向量间的相关度，以此作为文档间的相关度。余弦夹角是常用的相似函数之一。式（1）给出了利用这种

收稿日期：2014-03-05；修订日期：2014-05-08

基金项目：国家自然科学基金项目（61272109）

作者简介：侯超昆（1990-），男，河南新乡人，硕士研究生，研究方向为 Web 搜索与挖掘、数据挖掘；李石君（1964-），男，湖南岳阳人，博士，教授，CCF 会员，研究方向为 Web 搜索与挖掘、数据挖掘、大数据。E-mail：chaokunhou@qq.com

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方法计算文档 d_i，d_j 之间的相似度 $sim(d_i, d_j)$ 的公式

$$sim(d_j, d_k) = \frac{d_j \cdot d_k}{|d_j| \times |d_k|} = \frac{\sum_{i=1}^{n} w_{ji} \times w_{ki}}{\sqrt{\sum_{i=1}^{n} w_{ji}^2} \times \sqrt{\sum_{i=1}^{n} w_{ki}^2}} \tag{1}$$

向量空间模型具有强相似性的特征，且兼具简单高效等优点，已经在广泛的领域得到实际应用。但它没有从语义层次上考虑隐藏在特征词后面的概念之间的语义关系，这会造成同义词或多义词对相关度计算的干扰。同时特征词出现的位置也无法用向量表示。

还有一种在 VSM 的基础上被称为基于主题的向量空间模型（TVSM）的相关性分析模型[1]。TVSM 是将文档表示为一个基于所有词项的向量。有别于 VSM 的是，它将文档向量中的每一特征词项表示为一个 d 维向量，这个 d 维空间的每个维度代表一个基本主题，而向量的每个分量表示该特征词项与该主题之间的相关度。所有主题之间彼此独立。TVSM 同样是用文档向量之间的夹角余弦值表示文档之间的相似度。TVSM 在一定程度上能够避免同义词和多义词对计算的干扰，但它并没有给出用来确定每个词项的权重和词项间夹角的有效方法，这在词项空间规模庞大时局限明显。

综合考虑上述相关度计算模型的优点和局限性，本文提出的基于本体的网页主题相关度计算方法，将本体应用到 VSM 中特征词权重的计算中，以此弥补传统的 VSM 在语义层次上的欠缺。同时，通过定义并提取主题概念，避免了 TVSM 无法有效确定庞大词汇空间中词项权重的问题。

1.2　本体的应用

当前，本体已经在语义 Web、信息集成、信息检索以及人工智能等多个领域内得到广泛使用[2]。本体是一个源于哲学领域的词汇，用来表示理论哲学的基础和系统阐述现实世界中实体的状态。在计算机科学中，在理论上最为著名且普遍采用的是 Tom Gruber 在 1993 年对本体的描述[3]：本体是一种对于共享概念体系形式化的、具体而明确的说明。尽管表达方式各异，但大多数本体描述的都是个体、概念、属性与关系。领域本体是针对某个特定领域建立模型，它可以提供对一个领域可共享的、共同的理解[4]。如图 1 所示，是一个简单的对“计算机语言”这个概念描述的本体的局部示例。

作为领域知识描述和语义 Web 的支柱[5]，本体的应用可以弥补传统文本表示方法在语义层次上的欠缺。同时，将本体作为知识表达的载体具有可重用的优点[6]。因此，越来越多的研究者开始将本体应用在文本表示与概念间语义关系的计算中。文献［7］中基于 WordNet 本体计算特征项之间的距离，然后利用文档与特征项之间的互信息计算得到特征项的权值，以此作为文档向量中每一项的权值。

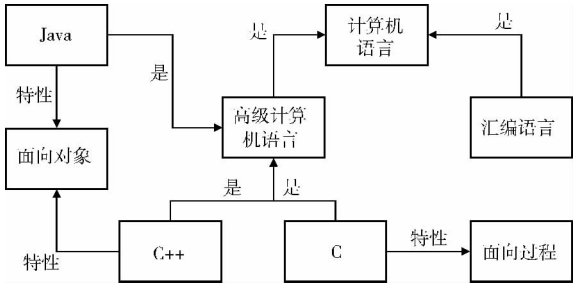

图 1　计算机语言概念本体局部

文献［8］给出了一种基于遗传算法的文本聚类方法，其中使用本体表示文本并计算本体中概念的语义相似度。在国内，2008 年，文献［9］通过利用本体能反映概念之间的语义相关度的特点，构建概念和文档的语义模糊集，用模糊集之间的相关度表示文本间的相关度。2010 年，文献［10］在文本聚类中引入本体，通过定义关键概念集，减少文本表示为向量的维度。李荣等人[11]在 2011 年提出了一种基于本体的旨在减小计算量的概念相似度计算方法。这种方法首先通过筛选得出候选概念集，然后针对该集合中的概念分别进行基于结构与基于属性的相似度计算。

结合上述研究，本文提出了一种基于领域本体的网页主题相关度计算模型 OBWTCCM（ontology based webpage-theme correlation calculation model）。

2　网页主题相关度计算

在 VSM 中，核心的问题是对表示文档的空间向量中各项进行赋值，这些值表示的是对应词汇在文档中的权重。已经有很多不同的方法计算这些权重值，其中最著名的是 tf-idf 算法。然而，这些方法归根到底都是一种基于词频统计的方法。现实文本中特征词的重要性还和与其有语义关系的其它特征词有关系。同时，不同的概念在不同领域中具有不同的重要程度。这些都导致了 VSM 对文本在语义层次上分析的不足。所以，首要问题就是对本体中概念之间的语义相关度进行计算并对领域特征词汇进行加权。OBWTCCM 是通过构建主题语义矩阵的方式来解决这个问题的。

2.1　构建主题语义矩阵

现实世界中，人们对事物的归类主要依据的是这件事物是否具有该类别所独有的特征。例如看到一篇文章，判断它是否属于某一领域，关键是看它是否包含有某一领域独有的信息。不同概念对一个主题的刻画程度是不同的。是否包含对主题具有良好指向性的概念往往能反映出文本与主题之间的相关程度。在此做如下定义：

定义 1　主题概念是指那些在某一领域中与主题高度相关的，并对主题具有较强指向性的特征词汇。

首要的工作是根据本体判断哪些概念可以作为主题概

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

念，并得到｛主题概念，权值｝的键值对。其中最简单的是人为确定主题概念，然而，这种方法往往缺乏客观统一的依据，对权值的判断也会受到个人因素的影响。同时，本体规模的增大意味着工作量的增加。下面给出一种根据本体概念间语义相关度提取主题概念并计算其权值的方法。

本体 O 可以用图 $G=(V, E)$ 来表示，V 是顶点集，用来表示概念，E 是边集，表示概念之间的语义关系。这种关系具有对称性、自反性和传递性。本体概念间的关系有同义、继承和属性。根据不同的关系，为概念节点之间的边赋予权重。其中同义关系表示两个概念在该领域内几乎没有差别，边权重最大，为 1；继承关系是上位概念与下位概念间为抽象与具体的关系，边权重次之；属性关系是下位概念对上位概念的描述，边权重较低。分别为其确定权值，如下所示

$$p=\begin{cases}1 & \text{同系}\\ 0.8 & \text{承系}\\ 0.6 & \text{性系}\end{cases} \tag{2}$$

此外，本体中，两个概念之间最短距离中边数越多，表示其语义间的差异越大。综合以上特点，本体中概念 c_i 与概念 c_j 之间的语义相关度 $sim(c_i, c_j)$ 的计算公式如下所示

$$sim(c_i,c_j)=\frac{\partial+weight\_dis(c_i,c_j)}{\partial+dis(c_i,c_j)} \tag{3}$$

式中：$dis(c_i, c_j)$ 表示本体的图 $G(O)$ 中代表 c_i 和 c_j 两个概念的顶点间边数最短的路径上边的条数，$weight\_dis(c_i, c_j)$ 表示这条最短路径上边的权重之和，∂ 是一个可调节系数。如果 $i=j$，则 $weight\_dis(c_i, c_j)=0$，$dis(c_i, c_j)=0$。这样，就可以得到每个概念与其它概念间综合相关度，作如下定义：

定义 2 本体中一个概念的语义影响力，指的是该概念与其它所有概念间的综合相关度。计算方法如下所示

$$t_i=\frac{a}{n}\sum_{k=1}^{n}sim(c_i,c_k) \tag{4}$$

式中：t_i——概念 c_i 的语义影响力，n——本体中概念的总数，a——一个可调节的系数。一个概念的语义影响力可以反映出该概念与本体所刻画的主题的相关程度。当这个值超过一定阈值时，可以认为这个概念具有对领域主题具有较强的指向性，即该概念可以作为这个领域主题的主题概念。这样就得到了主题概念集合 TC，如下所示

$$TC=\{c_i \mid c_i\in C, t_i\geqslant v\} \tag{5}$$

式中：C——本体中的概念集，t_i——概念 c_i 的语义影响力，v——设定的主题概念语义影响力阈值。

这样，根据式（6）就能得到主题语义相关度的矩阵 tc_matrix

$$\text{tc\_matrix}=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1n}\\ s_{21} & s_{22} & \cdots & s_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ s_{m1} & s_{m2} & \cdots & s_{mn}\end{bmatrix} \tag{6}$$

其中 $s_{ij}=sim(c_i, c_j)$，$c_i\in TC$，n 表示本体中概念的数量，m 表示主题概念的数量。接下来，就可以使用主题语义矩阵对领域主题和网页分别进行代数表示，构建网页主题相关度计算模型 OBWTCCM。

2.2 基于本体的网页主题相关度计算模型 OBWTCCM

OBWTCCM 是对向量空间模型 VSM 的一种改进。OBWTCCM 将网页与主题代数化表示为关于领域主题概念的向量，每个向量的项分别代表主题概念在网页与主题中的权重。以此将网页与主题间的相关度计算转换为两个向量之间夹角的运算。与 VSM 不同的是，在网页与主题的向量表示中，将基于词汇统计信息的 tf-idf 算法得到的特征词权重与特定领域的主题语义矩阵进行计算得到主题概念的权重，从而增加了特征词权重计算中语义层次的考虑。同时，向量表示只针对主题概念计算权重，有效降低了空间向量的维度，避免了高维度向量计算对效率的影响。

首先是主题的向量表示。在每一个网页中，它只包含了领域中的一个局部。对一个领域主题的整体特征进行描述，则需要属于该主题的一个网页集合。评估一个网页与特定主题的相关度，在一定程度上可以看作是对该网页与属于这个主题的网页集合的相关度计算。所以需要在领域专家的帮助下获得一个已明确属于该主题，且在分布上能够均匀覆盖该主题的所有概念的网页集合 TP 作为领域语料，它区别于未确定主题一般网页集合 P，两者之间的关系式 $TP\in P$。那么每个特征词的权重可依据 tf-idf 算法的思想得出。计算公式如下

$$w_i=tf_i\times idf_i=\frac{n_i}{N}\times\log\frac{|D|}{1+d_i} \tag{7}$$

这里需要说明，在进行主题向量表示和网页向量表示中，式（7）中变量含义有所不同。在主题的向量表示中，w_i 是特征词 c_i 在主题向量中的权重，tf_i 表示 c_i 在 TP 中的词频，idf_i 是 c_i 逆向文件频率，由 $\log\frac{|D|}{1+d_i}$ 得出，n_i 是 c_i 在 TP 中所有网页中的出现次数，N 是 TP 中所有网页文本中词的总数，$|D|$ 表示 P 中网页的总数，d_i 是 P 中有 c_i 出现的网页的总数。这样，就得到了所有特征词的权重向量 $W_1=\{w_1, w_2, \cdots, w_n\}^T$。根据式（8）可以将主题表示为向量 t_vector

$$t\_vector=tc\_matrix\times W_1=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1n}\\ s_{21} & s_{22} & \cdots & s_{2n}\\ \cdots & \cdots & \ddots & \cdots\\ s_{m1} & s_{m2} & \cdots & s_{mn}\end{bmatrix}\begin{bmatrix}w_1\\ w_2\\ \cdots\\ w_n\end{bmatrix}=\begin{bmatrix}tw_1\\ tw_2\\ \cdots\\ tw_m\end{bmatrix} \tag{8}$$

主题表示为向量后，接下来将目标网页 WP 表示为向量。其思想与主题的表示方法相同，区别在于在使用式（7）求每个特征词的权重时，n_i 表示的是 c_i 在该网页中的出现次数，N 表示该网页文本中词的总数。由此可得到该

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

网页的特征词权重向量 $W_2=\{w_1, w_2, \cdots, w_n\}^T$。将 W_2 替换式（8）中的 W_1，得到该网页的向量表示形式

$$p\_vector=[pw_1 \quad pw_2 \quad \cdots \quad pw_m]^T$$

计算向量 t_vector 与 p_vector 之间的夹角余弦值，以此作为目标网页与主题之间的相关度，公式如下所示

$$COR(T,WP_k)=\frac{t\_vector\times p\_vector_k}{|t\_vector|\times|p\_vector_k|}=\frac{\sum_{i=1}^{m}tw_i\times pw_i}{\sqrt{\sum_{i=1}^{m}tw_i^2}\times\sqrt{\sum_{i=1}^{m}pw_i^2}} \tag{9}$$

式中：$COR(T, WP_k)$ ——网页 k 与主题 T 之间的相关度，p_vector_k——网页 k 的向量表示形式，m 是主题表示向量 t_vector 和 p_vector_k 的维度，即领域本体中主题概念的个数。

2.3 算法描述

基于本体的网页主题相关度计算主要分为 3 个步骤，首先根据领域本体计算概念间的语义关系，提取主题概念并由此得到该领域的主题语义矩阵。然后利用网页集合根据 tf-idf 算法得到各个特征词基于统计信息的权重。最后将主题语义矩阵与特征词统计信息权重相结合，得到主题与网页的向量表示，计算两向量之间的夹角余弦值作为两者之间的相关度。算法描述如下：

输入：领域本体 O，领域主题网页集 TP，一般网页集 P，目标网页 WP。

输出：网页与领域主题相关度 COR。

（1）利用式（2）对领域本体 O 中概念结点间的边进行加权，然后由式（3）得到各个概念间的语义相关度 $sim(c_i, c_j)$。继而根据式（4）得到每个概念的语义影响力 t_i；

（2）设定主题概念影响力阈值 t，根据式（5）得到主题概念集合 TC；

（3）由式（6）得到该领域的主题语义矩阵 tc_matrix；

（4）使用式（7）得到每个特征词在主题网页集 TP 中的统计权重。继而根据式（8）将主题表示为向量 t_vector；

（5）使用式（7）得到每个特征词在目标网页 WP 中的统计权重。然后根据式（8）将 WP 表示为向量 p_vector；

（6）利用式（9）得到 WP 与主题间的相关度 COR。

在计算其它网页与本领域主题相关度的过程中，步骤（1）～步骤（4）不需要重复执行，循环执行步骤（5）、步骤（6）即可。

3 系统实现与结果分析

3.1 系统设计

本文提出的基于领域本体的网页主题相关度计算方法已经在某市地税局税收情报中心项目中得到应用。作为情报数据中心的数据源之一，互联网涉税信息在税收情报数据来源中有着不可或缺的作用。情报中心互联网信息包括以下六类税收主题：基础信息类、产业投资类、重点税源类、财产交易类、文化产业类和热点信息类。系统中构建了一个基于开源爬虫框架 Heritrix 的主题爬取工具，针对每一类，从指定税务相关网站上爬取网页并计算其与主题的相关度，作为进一步信息分析的基础。首先在税务领域专家的帮助下构建上述各个税收主题类的本体，构建本体的工具使用的是 protégé 4.3。网页爬取的信息源采用领域专家提供的与各个主题相关的站点。由于在正规网站中，相同内容类型的网页往往具有规律的 URL 结构。例如，网易新闻的 URL 结构往往是以 news.163.com 开头，其后再加上年份和日期的字样。所以根据需求为爬取工具添加一些 URL 过滤规则，可以指定爬取网站中特定类型的网页。然后对采集回来的网页进行相关度计算并存储网页与计算结果。系统中爬取工具架构如图 2 所示。

系统以 Java 语言实现，运行在由一台 DELL PowerEdge R715 架构服务器、一台 DELL PowerEdge R815 存储服务器和四台 DELL PowerEdge R815 能力服务器组成的安装部署有云计算管理软件的云计算平台上。网页是带有 HTML 标签的半结构化文档，在进行相关度计算前，需要对网页进行去标签处理。同时，网页文本中包含有大量导航、广告等非正文内容，为避免这些垃圾信息对网页主题判断的干扰，系统在网页分析模块中采用了“基于行块分布函数的通用网页正文抽取算法”抽取网页正文，该算法已经在许多应用中得到使用且具有很高的准确度。而在相关度计算中对网页文本的分词工具，采用的是具有良好性能的中科院 ICTCLAS 汉语分词系统。设定主题相关度阈值 $COR\_THR=0.25$。对于相关度大于阈值的网页，认为其与主题相关。通过收集并分析系统运行结果，检验基于本体的网页主题相关度计算模型 OBWTCCM 的在实际运行中的效果。

3.2 结果分析

本文从 2 个方面分析 OBWTCCM 的性能。首先是验证该方法在对与主题相关网页的发现中有效性。采用信息检索领域最常用到的检索效果指标召回率 R（recall）和准确率 P（precision），以及综合考虑召回率和准确率的 F 值。召回率 r=计算得到的与主题相关的网页数量 cp/实际与主题相关的网页数量 rp；准确率 p=计算得到的与主题相关的网页数中实际与主题相关的网页数量 rcp/计算得到的与主题相关的网页数量 cp；这里我们认为召回率与准确率在评定相关度计算方法性能中同等重要，所以取 $F=2RP/(R+P)$。如表 1 所示为根据系统运行结果结合用户反馈数据得到的 VSM、OBWTCCM 两种方法，在情报中心税收六类主题信息中上述指标下的比较。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

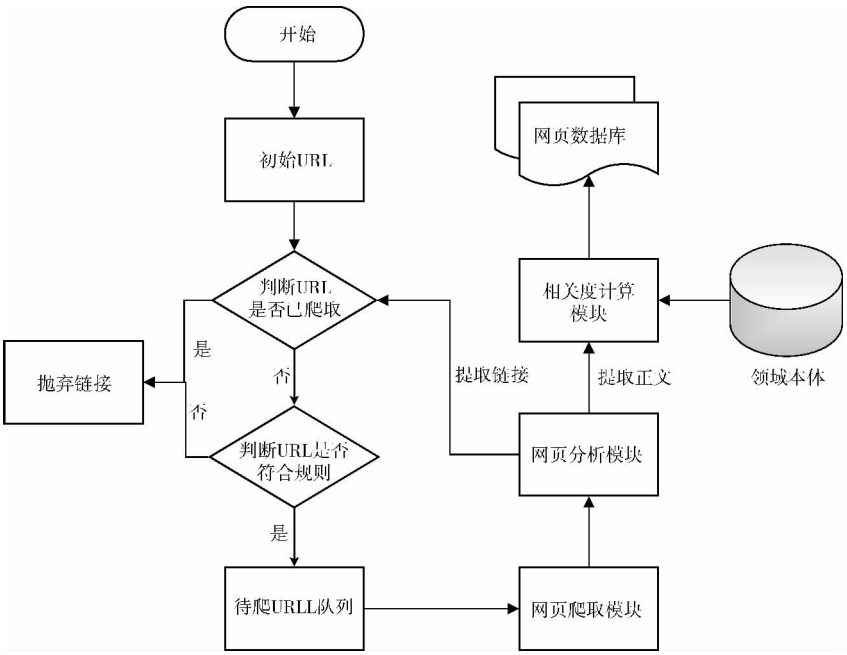

图 2 系统中爬取工具设计架构

表 1 两种方法对六类主题的效果比较

	VSM			OBWTCCM		
	R	P	F	R	P	F
基础信息	0.65	0.64	0.64	0.75	0.86	0.80
产业投资	0.64	0.67	0.65	0.82	0.83	0.82
重点税源	0.55	0.71	0.62	0.77	0.81	0.79
财产交易	0.61	0.72	0.66	0.79	0.85	0.82
文化产业	0.59	0.63	0.61	0.76	0.87	0.81
热点信息	0.57	0.69	0.62	0.82	0.91	0.86

由表 1 可知，OBWTCCM 在召回率、准确率和 F 值上相较 VSM 都有了明显的提高。这说明了 OBWTCCM 能够较准确地发现与主题相关的网页。接下来，针对网页个体，在更为精确的意义上检验该方法计算所得的相关度与实际情况之间的误差。

为避免与主题毫无相关的网页对结果的干扰，从 6 个主题中随机选取爬取回来的共 60 个与主题相关的网页。由 5 个领域专家按照事先约定的统一标准为每个网页与主题相关程度在 0 到 1 之间进行评定，取 5 个分数的平均值作为该网页与主题间相关度的专家评分。将其与 VSM 计算得到的相关度和本文 OBWTCCM 得到的相关度进行比较，比较结果如图 3、图 4 所示。

在图 3、图 4 中，网页 1-60 按照专家评分降序排列。专家评分可以近似看作实际相关度，VSM 和 OBWTCCM 两种方法计算所得的相关度曲线围绕专家评分曲线波动越小，表示方法计算得到的相关度在整体上越接近于实际相关度。比较可知，图 4 中 OBWTCCM 相关度曲线与专家评分曲线相较于图 3 中 VSM 相关度曲线在整体趋势上更为接近且波动幅度更小，这说明总体上由 OBWTCCM 计算所得的网页与主题之间的相关度与实际相关度之间的误差较小，具有较高的可信度。

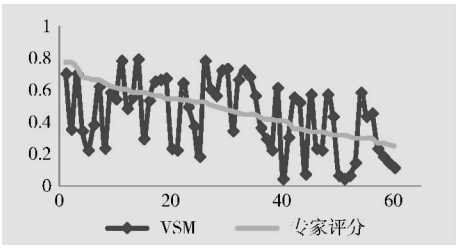

图 3 VSM 相关度与专家评分比较

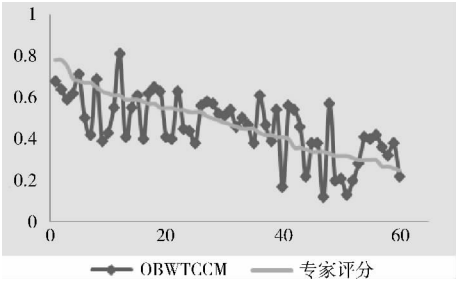

图 4 OBWTCCM 相关度与专家评分比较

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4 结束语

在对网页与特定主题之间相关度的计算中，传统 VSM 仅从特征词的统计信息上考虑计算相关度，忽略了隐藏在特征词之间的语义关系。本文提出了一种基于领域本体的网页主题相关度计算模型 OBWTCCM，通过计算本体概念间的语义相关度提取主题概念，结合特征词的统计信息为主题概念在主题中的权重进行赋值。在将主题与网页表示为向量进行相关度计算中，这种方法综合了特征词的统计信息与其之间的语义关系，弥补了 VSM 在语义层次上的欠缺。同时，该方法使用主题概念构造特征向量，降低了向量空间维度，避免了高维度向量空间计算造成的计算机负担。实际应用结果表明，该方法具有较高的有效性与准确度。下一步的工作将致力于垃圾网页的筛选和网页时效性对主题相关度计算影响的研究，以期进一步提高准确性。

参考文献：

[1] JIA Xiping, PENG Hong, ZHENG Qilun, et al. Topic-based document retrieval model [J]. Journal of South China University of Technology (Natural Science Edition), 2008, 36 (9): 37-42 (in Chinese). [贾西平，彭宏，郑启伦，等. 基于主题的文档检索模型 [J]. 华南理工大学学报（自然科学版），2008，36 (9)：37-42.]

[2] GAO Jianbo, ZHANG Baowen, CHEN Xiaohua. Research progress in security ontology [J]. Computer Science, 2012, 38 (8): 14-19 (in Chinese). [高建波，张保稳，陈晓桦. 安全本体研究进展 [J]. 计算机科学，2012，17 (9)：14-19.]

[3] HUANG Yinghui, LI Guanyu. Gruber's definition of ontology: Understanding in Chinese [J]. Computer Engineering and Design, 2008, 29 (8): 2125-2126 (in Chinese). [黄映辉，李冠宇. Ontology 的 Gruber 定义：中文语境理解 [J]. 计算机工程与设计，2008，29 (8)：2025-2126.]

[4] Lv Yanhui, Xie Chong. An ontology-based approach to build conceptual data model [C] //9th International IEEE Conference on Fuzzy Systems and Knowledge Discovery, 2012: 807-810.

[5] LI Jianghua, SHI Peng, HU Changjun. Overview of ontology search and ranking [J]. Journal of Chinese Computer Systems, 2013, 34 (10): 2396-2406 (in Chinese). [李江华，时鹏，胡长军. 本体搜索与排序方法研究综述 [J]. 小型微型计算机系统，2013，34 (10)：2396-2406.]

[6] ZHANG Zhiqiang, SONG Weitao, XIE Xiaoqin. An efficient ontology ranking algorithm—MIDSRank [J]. Journal of Computer Research and Development, 2011, 48 (6): 1077-1088 (in Chinese). [张志强，宋伟涛，谢晓芹. 一种有效的本体排序算法 MIDSRank [J]. 计算机研究与发展，2011，48 (6)：1077-1088.]

[7] Zhang Xiaodan, Jing Liping, Hu Xiaohua, et al. Medical document clustering using ontology-based term similarity measures [J]. International Journal of Data Warehousing & Mining, 2008, 4 (1): 62-73.

[8] SONG Wei, LI Chenghua, PARK S C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures [J]. Expert Systems with Applications, 2009, 36 (5): 9095-9104.

[9] SONG Ling, GUO Jiayi, ZHANG Dongmei, et al. Semantic similarity computation of concepts and documents [J]. Computer Engineering and Application, 2008, 44 (35): 163-167 (in Chinese). [宋玲，郭家义，张冬梅，等. 概念与文档的语义相似度计算 [J]. 计算机工程与应用，2008，44 (35)：163-167.]

[10] ZHU Huifeng, ZUO Wanli, HE Fengling, et al. A novel text clustering method based on ontology [J]. Journal of Jilin University (Science Edition), 2010, 48 (2): 277-283 (in Chinese). [朱会峰，左万利，赫枫龄，等. 一种基于本体的文本聚类方法 [J]. 吉林大学学报（理学版），2010，48 (2)：277-283.]

[11] LI Rong, YANG Dong, LIU Lei. Research of ontology-based conceptual similarity computation [J]. Journal of Computer Research and Development, 2011, 48 (Suppl.): 312-317 (in Chinese). [李荣，杨冬，刘磊. 基于本体的概念相似度计算方法研究 [J]. 计算机研究与发展，2011，48 (Suppl.)：312-317.]

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net