分阶段融合的文本语义相似度计算方法*

马军红

(西安外事学院工学院　西安 710077)

【摘要】面向中文文本的信息检索,提出一种从句子、段落到文本整体分阶段进行的文本相似度计算方法。该方法结合文档的主题与应用范围,用语义加强的权重计算方法对特征词赋予相应的权重,并根据每个计算阶段的特点,分别融入对文本语义的计算因素,力求使中文文本的相似度计算结果更为准确。最后建立文本相似度计算系统,通过与传统算法的实验结果进行对比,证明改进后的算法可以取得更好的效果。

【关键词】文本相似度　信息检索　语义相似度　权重

【分类号】TP391

A Staged and Integrated Semantic Similarity Algorithm of Text

Ma Junhong

(Engineering Institute, Xi'an International University, Xi'an 710077, China)

【Abstract】For Chinese text information retrieval, a staged and integrated similarity algorithm of text is proposed, which processes sentences, paragraphs and the whole document stage by stage. The algorithm combines the topic and application ranges of document, and the corresponding weight is given to the feature words via the weighted calculation method with the semantic enhancement. Moreover, these weights are integrated into the calculated factors of the text semantic with the characteristics of each calculation phase, respectively to reach the aim of finding a more accurate similarity calculation results for Chinese text similarity calculation. Finally, a text similarity computing system is built and the improved algorithm of the system achieves better experimental results comparing with the traditional algorithms.

【Keywords】Texts similarity　Information retrieval　Semantic similarity　Term weight

1　引　言

如今,各行各业的人们都能通过网络平台自由发布和下载信息,使得信息量不断增加,其中有大量重复和无用的信息。如何提高效率,轻松快捷地在这些信息中提取真正需要的东西,是信息处理领域的热点和难点。文本相似度的有效计算可以应用到文本分类、文本聚类、信息检索、问答系统、网页去重等很多领域。

文本信息是一种非结构化或半结构化的信息,它是现实生活中能获取的大部分信息的存在形式。就目前来说,尽管图像、视频等多媒体信息资源飞速增加,文本信息仍然占有相当大的比例,几乎达到70%以上[1]。然而,在文本相似度计算领域仍然存在不少问题需要人们解决,尤其是对中文文本相似度的研究。利用计算机来实现

收稿日期: 2013－07－05
收修改稿日期: 2013－09－02
* 本文系陕西省教育厅科学研究计划项目"基于实时嵌入式安全的双向序列加密方法研究"(项目编号: 2013JK1146)的研究成果之一。

自然语言理解，中文比英文更加困难。中文不像英文那样词与词之间有明显的分隔标记，它用多个连续的字词一起表达一个意思，根据上下文语境的不同，还容易引起歧义。缺乏坚实的理论依据和不能完全拟合文本的特性，是现有的文本相似度计算模型的弱点，如何改进中文文本的相似度计算方法，提高相关应用的效率，值得进一步研究和讨论。

2　文本相似度计算的特点及分析

2.1　文本相似度研究综述

文本相似度计算的核心是比较两个给定的文本（可以是字节流等）之间的差异，通常用［0，1］之间的一个数值来度量[2]。在不同领域，国内外学者已经提出了很多相似度计算方法并应用于实践。如向量空间模型（Vector Space Model，VSM）、布尔模型、隐含语义标引等统计模型、字符串匹配模型、基于语义理解的模型等。其中，VSM 是借鉴了统计学理论的向量概念而提出的。其核心思想是用向量来表示文本，将一个文本映射到 n 维的向量空间，可用 TF－IDF 方法来计算权重，建立文本的数学模型，这样文本间的相似度就可以用向量之间的关系来计算，最常用的是余弦距离计算方法[3]。字符串匹配模型是用字符串来代表文本的基本组成单位，如 LD 算法（Levenshtein Distance），又称为编辑距离算法（Edit Distance）[4]。这种算法是将字符串 A 通过插入字符、删除字符、替换字符变成另一个字符串 B，那么操作过程的次数就表示两个字符串的差异。基于语义理解的模型在进行相似度计算时，更多考虑的是文本的语义距离、语义相关性，希望深层次地挖掘文本含义，HowNet 知网与 WordNet 体系就是基于这种思想建立的[4]。但现在这类研究大多注重在词语和句子层面，取得的成果不多。

具体的说，文本相似度的计算在应用系统开发方面效果显著。主要有基于内容的搜索引擎，代表性的系统有北京大学天网、百度、慧聪等搜索引擎[5]。还有信息自动分类、自动摘要、信息过滤等文本级应用，如上海交通大学纳讯公司的自动摘要、复旦大学的文本分类、中国科学院计算技术研究所基于聚类粒度原理 VSM 的智多星中文文本分类器等[6]。可以看出，在信息处理各个领域，文本相似度的计算是无处不在的。

2.2　文本相似度的定义

相似度被应用于不同的领域，它所代表的意义也不一样。比如基于实例的机器翻译，一个词语将会用几个意思相同或相近的词语进行解释，这就侧重于考量词语之间的相似度；在 FAQ 自动问答系统中，用户查询的问句需要能迅速地在数据库中找到匹配问句的答案，这考量的是句子与句子之间的相似度，其中还要包含一定的语义相似度；在文本查重、版权维护与剽窃检测系统中，相似度则用段落与段落之间的相似度来度量[7]；在信息检索中，为了使检索结果更迅速、准确，需要对数据库中的文本集合进行分类、聚类、排序等操作，这就需要分析计算文本与文本之间的相似程度。

目前的研究主要关注查询词与文本的相似性，较少涉及文本与文本的相似性。本文主要研究的是整篇文本之间的相似度，如论文、申报书等。对于中文文本，在计算相似度时不应该单纯依靠词语的相近或者句子、段落的相似来判断，还应包含语义的理解，因此做如下定义：

文本相似度是指对于给定的两个或多个完整的文本，通过语句、段落的层层计算而得到它们之间的整体相似程度，同时包含了一定语义上的相似程度。

2.3　文本相似度计算的实现过程

计算中文文本的相似度，一般需要包括文本预处理、文本特征表示、特征选择和相似度计算几个过程。首先对输入的文本 D 进行文本预处理，在此基础上进行文本特征表示，然后经过特征选择抽取有用的特征，最后进行相似度计算，得到相似百分比[8]。将这个相似度计算结果与提前确定好的相似度阈值 Sim_{min} 进行对比，从而检测出相似的文档。

（1）文本预处理：是建立文本表示模型的基础，目的是将非结构的文本信息表示成结构化的形式。本文研究的是中文的文本相似度计算，这一部分的主要工作是中文自动分词和停用词处理。

（2）文本特征表示：一般是基于词的方法，即选取文本中意义较大的实词来表示，以双字词为主，也有三字词和四字词等多字词。除此之外还有 Shingling 算法，它是将相邻的两个或两个以上的词语以一个整体作为特征项。

（3）文本特征选择：在计算相似度时，希望待比较的意义不大的特征项能够减少，需要进行特征抽取和

降维处理。因此,文本特征选择成为文本相似性计算不可或缺的步骤。目前,文本特征选择可以使用权重的计算方法,或者定义某一个随机数。常用方法主要是:文档频率法(Document Frequency,DF)、信息增益法(Information Gain,IG)、互信息法(Mutual Information,MI)、$\chi 2$ 统计法(CHI)等[8]。其中 IG 和 CHI、MI 等属于有监督的特征选择算法,DF 属于无监督的算法,可以直接应用于聚类。

3 分阶段融合的语义相似度计算方法

本算法的基本思想是:分层次划分文本,将文本划分为段落,段落划分为句子,句子划分为词语;进行文本特征选择;分阶段计算词语、句子、段落的相似度,融合后计算出文本的相似度。在每个阶段融合语义相似度计算,完成局部到整体的结合。由于特征词的选择是非常重要的一个环节,为了提高最终的计算效率,本文先对传统特征词的权重算法进行改进。

3.1 语义加强的权重计算

在向量空间模型 VSM 中,文本被表示为由特征项组成的向量,而特征项的权重就是文本向量空间的坐标值[5]。特征词的权重对于文本的表示起着非常重要的作用,它的取值也影响着文本间相似度的计算结果。TF - IDF 权重计算方法是目前基于 VSM 使用最多的计算方法。但这种方法是基于词频的方法,如果一个词在文本预处理阶段被选择作为特征词,那么在计算词频的时候和其余被选择的词将不加区别同等对待。这种计算过于片面,没有考虑到特征词与文档内容相关性、上下文语境或者应用领域等其他因素。

因此,本文提出一种新的语义加强的权重计算方法,在前期文本预处理阶段计算词频时,可以结合文档的主题、范围、应用等,对所选中的特征词赋予相应的权重 f。

借助上述思想,本文将 TF - IDF 算法进行改进:

首先使用传统的 TF - IDF 公式来计算权重 w_{ij}:

设 L 为文本集合文本总数,w 代表权重,IDF 被定义为[10]:

$$IDF_t = \log_2 (L/DF_{ij}) \quad (1)$$

则用 TF - IDF 计算权重的公式如下[10]:

$$w_{ij} = TF_{ij} \times IDF_{ij} = TF_{ij} \times \log_2 (L/DF_{ij}) \quad (2)$$

然后从语义理解的角度考虑,得到新的计算公式:

$$w'_{ij} = \beta_1 \times w_{ij} + \beta_2 \times f_i \quad (3)$$

这里引入了两个参数 β_1 和 β_2,用传统公式计算的权重 w_{ij} 仍然占较大的比重,但需要结合语义影响的因素进行调整。β_1 和 β_2 两个参数之和为 1,其中 β_1 表示的是用基于词频统计的方法计算出来的权重值对最终权重值所占的比重。引入这个参数是由于基于词频统计的权重计算方法虽然有一定的科学性,但其提取的特征词不能完全表达出文本语义信息,例如“算法”与“方法”,一般同类别的科技性文章,通常都包含这两个词语,由于它们在整篇文章中可能出现的次数较多,会被优先选择为特征词条,而这些应该属于文本区分度较低的词语,不能简单因为出现的次数多而被设置为特征词。因此 β_1 的取值范围应在 0.6 - 0.9 之间。β_2 表示的是语义理解因素所占的比重,对应 β_1 的取值范围在 0.4 - 0.1 之间。这样可以根据文本的语义特征来计算权重,选取的特征词比简单依靠词频统计的方法所包含的信息量要高。而 f_i 的取值将根据特征词本身的属性来设置。一般文章的特征词分为三类:专业领域术语、非领域术语和一般词语。对于专业领域术语,其对于整篇文档的影响较大,因此 f_i 的取值为 1。非领域术语与一般词语的影响逐渐减低,其 f_i 值可设置为小于 1 的数。在具体实验时需要结合检测文本的类别、主题或关键词,在对应范围中取不同的参数值进行交叉对比。

综上,改进的权重计算方法加强了语义方面的考虑,更能体现特征项在文档中的影响程度。另外,新的计算公式中引入了三个参数,在计算时可根据需要具体设置,方法灵活实用。

3.2 文本语义相似度计算

通过分词系统将句子划分为若干词语,经过去除停用词等文本预处理操作,然后使用上述改进的 TF - IDF 算法提取特征词,接下来即可以分阶段地进行相似度计算。

(1) 词语相似度计算

借助知网的知识结构,在义原相似度计算时考虑其在义原分类树相对位置,利用如下公式[11]来计算:

$$WS(p1,p2) = \frac{a \cdot h_1 \cdot h_2}{d + a} \quad (4)$$

其中,p1、p2 表示两个义原,它们的语义距离为 d,代表 p1 和 p2 的路径长度,a 是一个可调节参数,它的值代表相似度为 0.5 时的路径长度。p1、p2 在义原树

中的深度，记作 h_1、h_2，代表它们在分类树上的相对位置，作为影响因素也考虑进去。

(2) 句子相似度计算

经过词语划分、特征抽取，句子 L(含 n 个特征词)和 R(含 m 个特征词)表示为以下向量形式:

$$L = (w_1, w_2, \cdots, w_n)$$

$$R = (w_1, w_2, \cdots, w_m)$$

可得到两个句子的相似特征矩阵 H(L,R):

$$\begin{bmatrix} S(L_1,R_1) & S(L_1,R_2) & \cdots & S(L_1,R_m) \\ S(L_2,R_1) & S(L_2,R_2) & \cdots & S(L_2,R_m) \\ \vdots & \vdots & & \vdots \\ S(L_n,R_1) & S(L_n,R_2) & \cdots & S(L_n,R_m) \end{bmatrix}$$

句子 L 与 R 的相似度计算公式为:

$$S(L,R) = \frac{\sum_{i=1}^{n} \max(S(L_i,R_1), S(L_i,R_2), \cdots, S(L_i,R_m))}{n} \quad (5)$$

即取矩阵中每行元素的最大值作为 L 的特征词 L_i 与 R 的相似度值，将这 n 个最大值相加，取平均值就得到了 L 与 R 的相似度。

为了能准确地得到这两个句子的相似度，本文采取的方法是利用公式(5)计算 S(L,R)与 S(R,L)，然后取两者的平均值，从而确保了计算结果的唯一性。

(3) 段落相似度计算

将段落看作是句子的集合，求其对应句子相似度的最大值。

设有两个段落 X、Y，X 被分为 t 个句子，Y 被分为 d 个句子，则相似度计算公式如下:

$$PS(X,Y) = \left(\frac{\sum_{i=1}^{t} \max(S(X_i,Y_1), S(X_i,Y_2), \cdots, S(X_i,Y_d))}{t} + \frac{\sum_{i=1}^{d} \max(S(Y_i,X_1), S(Y_i,X_2), \cdots, S(Y_i,X_t))}{d} \right) \times 0.5 \quad (6)$$

在算法设计时，对于段落中已经计算过相似度的句子组合，句子中已经计算过相似度的词语组合，均不再进行计算，直接赋予前值，以提高效率。

(4) 文本相似度计算

前面的计算使用的是取最大值相加再平均的方法，将段落中出现的词语、句子同等对待。这样对基于段落相似的文本整体相似度计算阶段来说过于简单和片面。本文考虑到根据段落所处文本的位置不同，其重要程度也就是对文章的影响也不同，应赋予不同的权重。如果关键段落相似，则整篇文本的相似度也随之提高。

设 D1，D2 为两个待比较的文本，D1 有 m 个段落，D2 有 n 个段落，为了方便起见，统一用 X 表示 D1 中的某个段落，用 Y 表示 D2 中的某个段落。则根据公式(6)计算出来的段落相似度记为: $PS_1(X_1, Y_j)$，$PS_2(X_2, Y_j)$，…，$PS_m(X_m, Y_j)$，j = 1,2,…,n。用 $SS_i(X,Y)$ 表示新的段落相似度值，则有:

$$SS_i(X_i,Y_j) = \beta_1 \times PS_i(X_i,Y_j) + \beta_2 \times w_i \quad (7)$$

其中，$\beta_1 + \beta_2$ 的值为 1，w_i 为段落的权重，具体取值均可根据需要设置。本文取 β_1 为 0.8，β_2 为 0.2，关键段落的权重 w_1 为 1，普通段落的权重 w_2 为 0.9。

则 D1 和 D2 的相似度计算公式如下:

$$Sim(D1,D2) = \frac{\sum_{i=1}^{m} SS_i(X_i,Y_j)}{m} \quad (8)$$

在实际的应用过程中，考虑到长文本通常段落比较多，很多较低的相似度对整篇文本影响几乎为零，可以预先设置一个相似度阈值，低于整个阈值的段落不再加入后面的权重计算，从而降低系统的开销，进一步提高效率。

4　文本相似度计算系统的组成与实现

4.1　系统的组成模块

系统包括了文本预处理(中文自动分词)、文本特征向量的表示、特征向量的选择和相似度的计算等步骤，主要由以下几个模块组成:

(1) 文本库模块: 用于存放要进行相似度计算的中文文本;

(2) 分词词典模块: 管理和维护作为中文分词依据的词典;

(3) 中文分词模块: 对给定的中文文本进行分词处理并进行歧义校正;

(4) 特征词抽取模块: 根据分词结果和词频统计分析，提取出代表待比较文本的特征向量及其中包含各词条对应的权值;

(5) 相似度计算模块: 计算文本特征向量的相似度。

4.2　系统的工作流程

(1) 从文本库中取出要进行相似度计算的两个文本;

(2) 使用中文分词器 IK Analyzer3.2 进行分词处理;

(3) 抽取文本特征向量并确定权值;

(4) 根据不同的相似度算法步骤计算文本之间的相似度。

文本相似度计算系统组成与流程如图 1 所示:

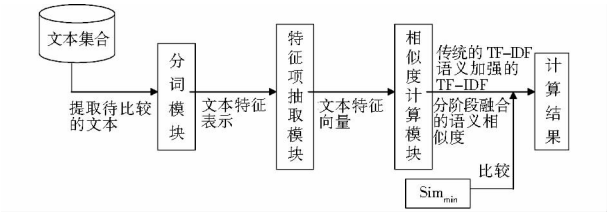

图 1 文本相似度计算系统组成与流程

5 实验与结果分析

5.1 实验数据集

本实验采用某研究单位的测试数据集,使用 MySQL 数据库。从中选取了 210 篇计算机、汉语语言文学、经济管理方向的文本,根据不同的篇幅分为三组,使用三种相似度计算方法分别进行测试,即:

(1) 传统 TF - IDF 方法: 传统 TF - IDF 方法是基于 VSM 的,利用余弦函数计算相似度;

(2) 语义加强的 TF - IDF 方法;

(3) 分阶段融合的语义相似度计算方法。

为了更好地验证本文提出的改进算法,先将传统的 TF - IDF 权重计算方法和改进后的 TF - IDF 方法进行实验对比,然后再使用上述三种方法进行完整的相似度计算实验。

实验选取的文本篇幅分布情况如表 1 所示:

表 1 实验文本篇幅分布

分组编号	文本篇幅范围(字符数)	文件数目
1	1 000 - 3 000	70
2	3 000 - 6 000	70
3	6 000 - 10 000	70

评估相似度计算的效果一般使用的评估指标是 F - measure 值,简称 F 值。F 值是准确率(Precision, P)与召回率(Recall, R)的协调均匀数。F 值的计算公式如下[13]:

$$F = \frac{2PR}{P+R} \tag{9}$$

其中:

$$P = \frac{正确检测到的相似文本数}{所有检测到的相似文本数}$$

$$R = \frac{正确检测到的相似文本数}{实际存在的相似文本数}$$

如果 F 值与 1 越接近,说明 Precision 和 Recall 均衡得越好。相反,如果 F 值与 0 越接近,则表示两个参数均衡性越差。

5.2 实验结果与分析

(1) 传统 TF - IDF 权重算法与语义加强的 TF - IDF 算法

在这两种算法中,均使用不同的相似度阈值进行实验对比,大于等于此阈值的被视为相似文本[7]。实验对比结果如图 2 所示:

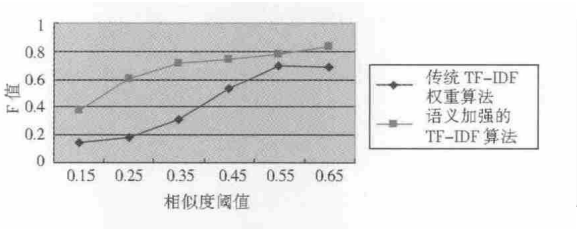

图 2 两种算法 F 值对比

从图 2 可以看出,语义加强的 TF - IDF 算法的效果总体比传统的 TF - IDF 权重计算方法效果要好。对于 0.25 - 0.55 之间的相似度阈值,其取得了比较稳定的 F 值变化,可靠性较强,而这个阈值范围正是在检测相似度的时候最常用的范围。而基于 TF - IDF 算法的 F 值显示其均衡性较差,最高值也没有达到 0.7,尤其在阈值为 0.15 - 0.55 之间结果很不稳定。这主要是因为传统的统计算法使得一部分低区分度的词成为了特征词,算法的准确率不高。而添加了语义因素的算法,考虑了文本的属性、类别等影响,更能体现文本的特性,计算的结果也较为准确。

(2) 不同相似度计算方法对比

经过上述实验比较,可以得出添加语义因素的算法是可行的。本文将分阶段融合的语义方法与前两种方法应用于相似度计算中。由于实验数据较多,现取相似度阈值为 0.3 为例。第 1 组实验结果如表 2 与图 3 所示:

表 2 相似度计算方法的对比 1

评估方法	传统的 TF - IDF	语义加强的 TF - IDF	分阶段融合的语义相似度
P 值	0.721 3	0.697 5	0.860 9
R 值	0.558 1	0.756 2	0.881 4
F 值	0.629 3	0.725 7	0.871 0

第 2 组实验结果如表 3 与图 4 所示。

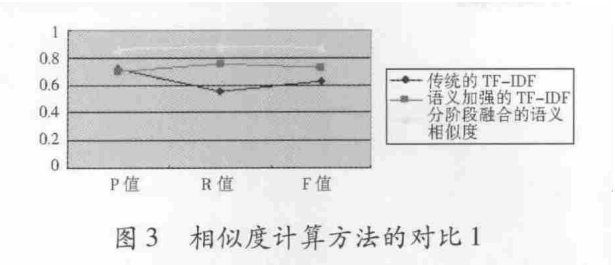

图 3　相似度计算方法的对比 1

表 3　相似度计算方法的对比 2

评估方法	传统的 TF – IDF	语义加强的 TF – IDF	分阶段融合的语义相似度
P 值	0.668 2	0.713 3	0.800 1
R 值	0.549 6	0.742 1	0.783 9
F 值	0.603 1	0.727 4	0.791 9

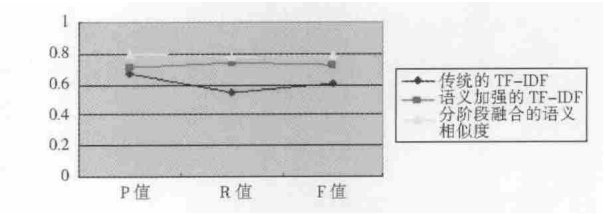

图 4　相似度计算方法的对比 2

第 3 组实验结果如表 4 与图 5 所示:

表 4　相似度计算方法的对比 3

评估方法	传统的 TF – IDF	语义加强的 TF – IDF	分阶段融合的语义相似度
P 值	0.683 4	0.774 2	0.843 1
R 值	0.661 5	0.769 4	0.838 6
F 值	0.672 3	0.771 8	0.840 8

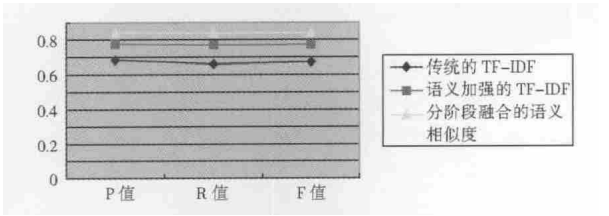

图 5　相似度计算方法的对比 3

根据准确率和召回率的含义分析可知,当准确率很高但是召回率很低时,说明该系统查找相似文本的命中率很高,也就是误判的数量少,但是查找不全面,即还存在许多相似文本没有找出;相反,如果召回率较高而准确率不高,就是把许多非相似的文本看成是相似的,即引起了很多误判,但是实际存在的相似文本大部分都已找出[12]。这两种情况皆是不理想的,只有二者相当,才能说明找出来的相似文本确实相似,并且把实际存在的相似文本已经基本上都找出来。通过 F 值调和平均,可以对相似度计算方法的效果进行均衡评估。从上述三组实验数据可以看出,本文的分阶段融合的语义相似度计算方法的准确率和召回率基本相当,F 值显示这两个参数的平衡性也较好,比其他方法的相似度计算准确性得到了提高。其中语义加强的 TF – IDF 方法也取得了较好的效果,传统基于 VSM 的 TF – IDF 方法运行不太稳定,效率较低。这个结果体现出对于中文文本的相似度计算,语义理解的方法要比单纯的统计计算方法更适合。

6　结　语

目前,越来越多的专家学者研究文本相似度计算,这是因为文本相似度的有效计算可以起到提高检索效率、避免文章剽窃、节省存储空间等作用。而中文的文本相似度计算处理非常复杂,在具体应用中还有很多不确定性。本文从语义理解的角度出发,对传统算法做进一步的改进,使得中文文本相似度计算的效率有所提高。在后续的实验中还可以采用其他方法进行对比,并加大实验数据规模和实验次数,更准确地分析和发现本方法的优缺点,使之运行更加稳定。另一方面,如何建立一个更好的语义理解模型,把它应用到更多的具体领域进行验证,比如文本检索、信息查询、文本聚类等,也是下一步研究的重点。

参考文献:

[1] 赵辉,刘怀亮,范云杰. 复杂网络理论在中文文本特征选择中的应用研究[J]. *现代图书情报技术*,2012(9):23 – 28. (Zhao Hui, Liu Huailiang, Fan Yunjie. Study on the Application of Complex Network Theory in Chinese Text Feature Selection [J]. *New Technology of Library and Information Service*,2012(9):23 – 28.)

[2] 金希茜. 基于语义相似度的中文文本相似度算法研究[D]. 杭州:浙江工业大学,2009. (Jin Xiqian. Chinese Text Similarity Algorithm Research Based on Semantic Similarity [D]. Hangzhou: Zhejiang University of Technology, 2009.)

[3] 舒晓明. 基于语义网的个性化信息检索的研究与实现[D]. 沈阳:沈阳工业大学,2011. (Shu Xiaoming. Research and Realization of Personalized Information Retrieval Based on Semantic Web [D]. Shenyang: Shenyang University of Technology, 2011.)

[4] 陈涛,林杰. 基于搜索引擎关注度的网络舆情时空演化比较分析——以谷歌趋势和百度指数比较为例[J]. *情报杂志*,2013,32(3):7 – 11. (Chen Tao, Lin Jie. Comparative Analysis of Temporal – Spatial Evolution of Online Public Opinion Based on Search Engine Attention——Cases of Google Trends and Baidu Index [J]. *Journal of Intelligence*,2013,32(3): 7 – 11.)

[5] 王静帆. 基于文本相似度的二阶段招聘信息检索[D]. 北京:

清华大学，2007.（Wang Jingfan. Two－Step Job Information Retrieval Based on Document Similarity [D]. Beijing: Tsinghua University，2007.）

[6] 谭慧琳，刘先锋. 基于遗传算法的知识推理研究[J]. 电脑知识与技术，2011，7(31)：55－59.（Tan Huilin，Liu Xianfeng. The Research of the Selection of Knowledge Reasoning Method Based on Genetic Algorithm [J]. *Computer Knowledge and Technology*，2011，7(31)：55－59.）

[7] 路永和，李焰锋. 多因素影响的特征选择方法[J]. 现代图书情报技术，2013(5)：34－39.（Lu Yonghe，Li Yanfeng. A Feature Selection Based on Consideration of Multiple Factors [J]. *New Technology of Library and Information Service*，2013(5)：34－39.）

[8] 黎邦群. 基于 Mashup 的特殊词快捷检索及检索建议[J]. 图书情报工作，2012，56(17)：126－130.（Li Bangqun. Quick Search of Special Words and Search Suggestions Based on Mashup [J]. *Library and Information Service*，2012，56(17)：126－130.）

[9] Duan Y X，Lei H. The Formal Definitions of Semantic Web Services and Satisfiability [J]. *International Journal of Advancements in Computing Technology*，2012，4(23)：327－335.

[10] Lee M C. A Novel Sentence Similarity Measure for Semantic－based Expert Systems [J]. *Expert Systems with Applications*，2011，38(5)：6392－6399.

[11] 王蕊，冯登国，杨轶. 基于语义的恶意代码行为特征提取及检测方法[J]. 软件学报，2012，23(2)：378－393.（Wang Rui，Feng Dengguo，Yang Yi. Semantics－based Malware Behavior Signature Extraction and Detection Method [J]. *Journal of Software*，2012，23(2)：378－393.）

[12] 刘兵. Web 数据挖掘[M]. 北京：清华大学出版社，2011：113－119.（Liu Bing. Web Data Mining [M]. Beijing：Tsinghua University Press，2011：113－119.）

（作者 E－mail：maxiaofei913@163.com）

Kuali 基金会和 EBSCO 合作改善 Kuali OLE 发现功能

Kuali 基金会很高兴 EBSCO 加入 Kuali 商务合作伙伴社区。EBSCO 将为 Kuali OLE（开放图书馆环境）提供其专业的无缝集成发现服务，以增强印刷资源和电子资源的访问。EBSCO 作为全球的领导者，提供在线研究数据库和发现服务，为世界各地的企业、高校、图书馆及其他机构提供服务。

“EBSCO 是一个值得欢迎的商务合作伙伴”，Kuali OLE 委员会联合主席 Deborah Jakubs 表示“EBSCO 能为我们带来有关研究型图书馆发现服务解决方案的扎实的专业知识，并且，EBSCO 已经和 OLE 就促进图书馆业务系统和结构化数据及发现服务的互操作进行过合作。我们很期待与 EBSCO 再一次进行合作，为改善 OLE 社区的发现服务共同努力。”

EBSCO 和 Kuali OLE 的这一合作目的是为 Kuali OLE 社区的发现服务提供灵活的整合策略。EBSCO 首席信息官 Michael Gorrel 说“EBSCO 和 Kuali 的合作旨在为图书馆提供独特的发现服务体验。OLE 所基于的开放标准对实施了 OLE 的图书馆和机构来说是一个很诱人的替代物。这也表示，顶尖大学已准备将学术图书馆驶向新的方向。作为一个发现服务提供商，我们希望为客户提供更多的服务，帮助图书馆更好地改善发现体验，更好地将内容和管理进行结合，以满足图书馆员和终端读者用户的需求。”

（编译自：http://www.librarytechnology.org/ltg－displaytext.pl?RC=18084）

（本刊讯）