

信息检索中相似度计算的新方法

陈 琴

(广西大学 计算机与电子信息学院, 广西 南宁 530004)

摘 要: 提出一种计算相似度的新方法, 在计算相似度时考虑到查询词出现在文档中不同部分对相似度计算的不同贡献, 即不同的权重系数, 并用遗传算法确定各个权重系数. 实验结果表明, 这种新方法在信息检索中在计算相似度时比传统的向量空间模型方法具有更好的效果.

关键词: 相似度计算; 权重系数; 遗传算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)09-0159-03

New Approach to Similarity Computing in Information Retrieval

CHEN Qin

(School of Computer and Electronics Information, Guangxi University, Nanning 530004, China)

Abstract: Similarity computing plays an important role in information retrieval. In this paper, a new approach to computing similarity is proposed. When we compute similarity, different contribution, namely weighted coefficients of different position of words in different part of the document are considered. And we use Genetic Algorithm to consider the weighted coefficients. Experimental results show that the new method has more effect than Vector Space Model.

Key words: similarity computing; weighted coefficients; genetic algorithm

1 引言

相似度计算研究的是用什么样的方法来计算或比较两个对象的相似程度. 在信息检索领域, 相似度的计算结果直接决定了检出对象的排名顺序. 面对搜索引擎返回的大量结果, 用户往往无法全部浏览, 所以对象的排名顺序尤为重要^[1]. 在信息检索中, 计算相似度的传统方法有向量空间模型^[2], 它把文档和查询词表示成以词为元素单位的向量, 根据词频 tf 以及逆文档频率 idf, 赋予该向量各个分量的权值, 与欧氏空间的向量一一对应, 用向量夹角的余弦值定量表示文档和查询词之间的相似度^[3,4], 即

$$q = (w_{q1}, w_{q2}, \dots, w_{qn})^T \quad (1)$$

$$d = (w_{d1}, w_{d2}, \dots, w_{dn})^T \quad (2)$$

$$\text{sim}(q, d) = \frac{\sum_{i=1}^n w_{qi} \cdot w_{di}}{\sqrt{\sum_{i=1}^n w_{qi}^2} \cdot \sqrt{\sum_{i=1}^n w_{di}^2}} \quad (3)$$

用户在进行检索时, 系统根据用户输入的查询词计算出每个文档与查询词的相似度, 进而根据相似度的大小, 将查询的结果进行排序并按照排序的先后返回给用户.

向量空间模型可以实现文档的自动分类和对查询结果的相似度排序, 能够有效提高检索效率; 它的缺点是相似度的计算量大, 当有新文档加入时, 则必须重新计算词的权值.

2 词语相似度计算的新方法

2.1 对文档的处理

进行信息检索时, 在看到全文之前大多有一级过渡页面, 其中包括了标题、作者、摘要、关键词等各种元数据^[5]. 因此, 可以认为, 一篇文档由标题、摘要、关键词、正文几个部分组成, 而每一部分的作用是不一样的, 也就是说, 各个部分对文档的贡献是不一样的. 一般而言, 标题对文档的概括能力最强, 关键词一般能代表文档在专业方向上的一些重要词

汇,而摘要部分则是对文档内容的一个较为详细的描述.对于用户的查询词,如果文档的标题出现了该查询词,则该文档符合用户要求的可能性是很大的.为此,我们在计算相似度时引进了文档不同部分的不同贡献,即不同的权重系数.

在计算相似度前首先对文档进行处理使之与查询词相联系,方法是确定用户输入的查询词及其同义词在文档中出现与否及出现的位置,这里除了确定查询词在文档中出现情况,同时还确定查询词的同义词在文档中出现情况^[6].接下来,依次按照标题、摘要、关键词、正文的顺序来匹配查询词及其同义词并做如下约定:(1)如果查询词出现在文档的某个部分,则用1来表示;(2)如果其同义词出现在文档的某个部分,则用0.6来表示,这里不用1表示,主要是考虑作为查询词的同义词,它只是在意义上也极有可能符合用户,并不能完全与查询词等同,因此,用小于1的数来表示,这个数值也可以根据实际情况取其他的值;(3)如果文档中的任何位置都没有出现查询词或其同义词,则用0表示.按照上面的约定,一个文档可以用一个4位的码字与用户的查询词相联系,这个4位的码字的形式为: $D_i = (X_1 X_2 X_3 X_4)$,比如,用户输入的查询词为“计算机”,若文档 D 的码字为(1001),表示查询词“计算机”在文档 D 的标题和正文中出现,而没有在摘要和关键词中出现.

2.2 相似度计算

在计算相似度前首先要确定文档集对应的码字集,然后按式(4)计算文档 D 与查询词 R 之间的相似度:

$$\text{sim}(D, R) = \frac{1}{4} \sum_{i=1}^4 \alpha_i x_i \quad (4)$$

其中, α_i 表示文档不同部分的权重系数, x_i 表示文档码字中的第 i 个分量的值.只考虑文档的标题、摘要、关键词和正文四个部分,因此 i 的取值为1~4.对于 α_i ,因为它代表不同部分对查询结果的贡献程度,取值如何直接影响到相似度的计算结果,从理论上讲,取0~1之间任意数值均可以,但这样取值并不能很好地反映不同部分对相似度的影响程度.因此,如何更好地确定 α_i 是相似度计算结果优劣的关键.文中利用遗传算法来确定 α_i 的取值.

用式(4)来刻画文档相似度是合理的,首先它作为相似度的值介于0和1之间,当文档 D 的4个部分都不包含查询词 R 时, $\text{sim}(D, R) = 0$,表明文档 D 与查询词 R 之间不相似,因此,文档 D 不应该作

为查询结果返回给查询用户;如果文档的若干部分包含查询词 R ,则 $\text{sim}(D, R) \neq 0$,表明文档 D 与查询词 R 之间相似, $\text{sim}(D, R)$ 的值越大,表明文档 D 与查询词 R 之间越相似,返回给查询用户时它应排在越靠前的位置.由此可以看出,利用式(4)来计算相似度确实反映了文档与查询词之间的相似程度.

3 利用遗传算法确定权重系数

遗传算法(genetic algorithm, GA)是基于自然选择和遗传学原理的随机搜索算法.由于遗传算法在解决非线性问题中表现出较强的鲁棒性、全局优化性和并行处理等优点,因此在复杂函数优化、生产调度、图像识别、机器学习等众多领域都获得了成功的应用.遗传算法的基本框架是编码、初始种群、适应度函数的设计、选择操作、交叉操作和变异操作^[7].文中权重系数的确定,实质上可以看成是在其取值范围内搜索最优的系数组合,以达到最优的计算效果.因此,利用遗传算法来确定权重系数从理论上的可行的.

利用GA确定权重系数的步骤:

(1) 编码

把待定的四个权重系数表示成一维数组 $x = (\alpha, \beta, \gamma, \delta)$,数组各元为双精度实数.

(2) 初始群体的随机生成

设群体规模为 n ,随机生成初始群体的 n 个个体 $x_i = (\alpha_i, \beta_i, \gamma_i, \delta_i), i = 1, 2, \dots, n$.

(3) 适应度函数的设计

待定权重系数的确定应该使得查询返回的结果与用户想要得到的结果的吻合程度最好,因为不同的用户返回结果的要求不尽相同,目前也没有一个统一的评价方法,文中采用以专家的评价作为评价标准,即对于用户输入的查询词,返回的文档及其返回排名由专家来确定,利用系统返回的文档排名与专家的排名之间的差距作为衡量返回结果的优劣.因此,使用以下公式^[8]作为适应度函数:

$$D(R', R) = \frac{\sum_{i=1}^n [(n-i) \times \sum_{j=1 \wedge R'_j \in \{R_1, \dots, R_i\}} 1]}{\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} [(n-i) \times i] + \sum_{i=\lfloor \frac{n}{2} \rfloor+1}^n (n-i)^2} \quad (5)$$

其中, n 是专家排名列表中的文档数, R'_j 代表系统返回的第 i 篇文档, R_i 代表在专家给出的排名列表 R 中的第 j 篇文档.可以看出,公式的分子部分是用

来衡量两个排名列表的实际距离,而分母则是用来把距离归一化的.对于排名在前的文献的错排,此公式给予了更大的“惩罚”.例如:如果最好的文献错排了,对于错误的权重是 $n-1$,而对于第二好的文献错排,对于错误的权重则是 $n-2$.所以 D 的值说明在权重系数组合下系统返回的排名与专家排名差距.当然,这个差距越小,表明系统返回的结果与用户想要得到的结果的吻合程度越好.

(4) 选择操作:利用锦标赛选择(tournament selection)方法进行选择,选择适应度小的个体作为优良个体,对其复制,淘汰适应度大的个体.

(5) 交叉操作:对配对池 $MP = (x_1, x_2, \dots, x_i, \dots, x_j, \dots, x_n)$ 中的 n 个个体进行交叉,若选择 $x_i = (\alpha_i, \beta_i, \gamma_i, \delta_i)$ 与 $x_j = (\alpha_j, \beta_j, \gamma_j, \delta_j)$ 配对,则交叉操作为

$$x'_i = (\alpha_p, \beta_q, \gamma_t, \delta_q)$$

$$x'_j = (\alpha_q, \beta_p, \gamma_p, \delta_t)$$

其中,交叉参数 p, q, t 为随机数.

(6) 变异操作:若选择 $x_i = (\alpha_i, \beta_i, \gamma_i, \delta_i)$ 进行变异,则变异的个体为 $x'_i = (\alpha_p, \beta_q, \gamma_t, \delta_s)$,其中 p, q, t, s 为随机数,范围为 $0 < p, q, t, s < 0.33$.

4 实验

从中国期刊全文数据库上收集了 900 篇科学文献,这些文献涉及计算机、农业、经济管理等 4 个领域,按照不同的领域将数据分为 4 组,并使用每组的 2/3 的数据作为训练集来确定权重系数,另外的 1/3 的数据作为测试集,分别用传统的向量空间模型方法和文中方法来计算相似度,使用与专家排名列表之间的距离差对结果进行评估,实验结果如表 1 所示.

表 1 传统向量空间模型与本文方法性能比较

组号	传统的向量空间模型方法	文中方法
1	0.598	0.312
2	0.781	0.405
3	0.526	0.301
4	0.686	0.423

实验结果表明利用文中的相似度的计算方法比传统的向量空间模型方法的性能有较大提高.

5 结束语

文中提出了相似度计算的一种新方法.与其他相似度计算公式相比较,该方法考虑了查询词出现在文档的不同部分对查询结果的影响程度不同,并采用码字的方法来表征文档信息的特征.该方法计算简单,完全避免了诸如在欧氏空间中求相似度的大量乘法运算,可较大地提高计算速度.实验表明,该方法是有优越性的.

参考文献:

- [1] 张纯青,陈超.基于加权词频的信息检索相似度评价模型[J].计算机仿真,2008,25(1):134-137.
- [2] Salton G, Lesk M E. Computer evaluation of indexing and text processing[J]. Journal of the ACM, 1968, 15(1):8-36.
- [3] Salton G. Automatic text processing: the transformation analysis and retrieval of information by computer [M]. Pennsylvania: Addison-wesley, 1989.
- [4] Sheridan P, Ballerini J P. Experiments in multilingual information retrieval using the SPIDER system [C]//Proc the 8 Annual A CMSIGIR Intl. Conf. on Research and Development in Information Reterieval. Canada, 1996: 58-65.
- [5] 钱建立,吴广茂,蒋路.基于特征相似度的科技论文元数据提取算法研究[J].微电子学与计算机,2008,25(8):129-132.
- [6] 陈沈焰,吴军华.基于本体的概念语义相似度计算及其应用[J].微电子学与计算机,2008,25(12):97-99.
- [7] 赵金帅.一种用于防止早熟收敛的改进遗传算法[J].西南大学学报:自然科学版,2008,30(1):156-159.
- [8] Zaiqing Nie, Yuanzhi Zhang, JiRong Wen. Object-level ranking: bringing order to web objects [C]//International World Wide Web Conference. USA: Hawaii, 2005: 567-574.

作者简介:

陈 琴 女,(1974-),硕士,讲师.研究方向为信息检索、电子商务.