语义检索系统中的查询语句扩展算法改进

杨学兵[1,2],钱 蓉[2]
(1.南京大学 计算机科学与技术系,江苏 南京 210093;
2.安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:查询扩展技术是在原有用户查询的基础上加入语义相关的新词,组成语义更准确的查询条件。文中对查询扩展算法中扩展词加权方法进行改进,提出一种基于初始用户查询意欲和词与词间语义关联性给扩展词加权的方法。根据此算法得到的扩展词权值不仅反映了该扩展词和原关键词间的关联性,还反映出该扩展词和查询关键词集合中所有元素的关联性。因此,可将基于语义树的查询扩展问题转换为扩展词权值 $w^{ij}_{s,o,p}$ 的计算,如何计算出权值 $w^{ij}_{s,o,p}$ 是文中的核心。实验证明,该算法提高了检索的查准率。

关键词:语义相似度;语义关联度;查询扩展;语义检索
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2008)12-0001-03

Improvement on Arithmetic of Query Expansion in Semantic Retrieval

YANG Xue-bing[1,2], QIAN Rong[2]
(1.Dept.of Computer Sci.and Techn.,Nanjing University,Nanjing 210093,China;
2.Sch.of Computer,Anhui University of Technology, Maanshan 243002,China)

Abstract: Query expansion adds some new correlated words to original retrieval to make a more efficient query. In order to improve the way of adding the weight to the word in the query expansion, proposes a new way of adding the weight based on original query and the relation between a word and another. According to this arithmetic, it can get the weight which reflects not only the original word and expansive word but also the original query and expansive word. Therefore, the problem of query expansion can be converted into computing the value of "$w^{ij}_{s,o,p}$", the core of this paper is how to compute the value of $w^{ij}_{s,o,p}$. The experimental result shows this arithmetic improve the accuracy of semantic retrieval.

Key words: semantic similarity; semantic relation; query expansion; semantic retrieval

0 引 言

随着Internet信息技术的发展,网络上的信息量正在飞快增长,导致用户很难在海量信息中发现自己所需的信息。因此,如何在Web信息中快速检索到需要的信息已成为人们关注点的焦点,也是众多学者研究的热点。

1 背景及相关研究工作

在当前的信息检索模型与系统中,用户的查询请求通常以关键词的形式出现,传统信息检索利用简单的词匹配法则计算文档特征值与检索词间的相似度,往往只有查询词出现在文档中才能检索到。在自然语言中存在同义词和同音词的问题以及用户不能准确描述的问题,致使词不匹配成为影响信息检索效果的重要原因之一。

学者Van Rijsbergen[1]在1986年指出了仅限于原查询词来提高系统的检索性能是有限的,必须对原查询进行修改以提高检索性能。普遍认为Van Rijsbergen提出的对原查询的修改即查询扩展,主要涉及原查询关键词的权重修改和加入与原查询词相关的词。然而传统的查询扩展虽然在技术上有了很大改进,却不能实质性地提高信息检索性能,主要原因是传统的查询扩展技术是以查询词为中心,从机械式匹配层次上进行的查询扩展,忽略了查询语义及查询概念语义之间关联扩展,因而没有充分表达和扩展用户查询意图,不能从根本上消除用户查询意图与检索结果之间的语义偏差和用户查询的歧义性问题,也就没有最终解决查全率和查准率问题。

针对这些问题,近几年内有关学者开始作新的研

收稿日期:2008-03-26
基金项目:安徽省自然科学基金重点资助项目(2004KJ053ZD)
作者简介:杨学兵(1967-),男,安徽巢湖人,教授,研究方向为数据挖掘。

究,最受关注的是语义概念查询扩展[2]技术的研究,试图从语义概念的层次上扩展用户查询,取得了积极的成果。语义概念查询扩展必须实现同义词扩展、语义蕴涵扩展、语义外延扩展和语义相关扩展。语义蕴涵扩展是通过概念之间的语义蕴涵关系实现;语义外延扩展是通过概念之间的语义外延关系实现;而语义相关扩展是通过概念之间的语义相关关系实现,概念间的因果关系、特征关系、相互作用关系、对应关系等都可能被理解为概念的语义相关关系。

语义概念查询扩展的过程是首先要建立概念语义空间,然后从概念语义空间中提取用户查询语义及其语义关联,实现语义概念扩展。近几年来语义概念查询扩展已经成为研究热点,目前重要的研究方法是根据概念空间中概念间的相互关系,利用相关的技术把查询语义关系看作几个概念,从语义概念空间中提取出查询语义、语义相似概念以及语义相关概念,实现查询概念的扩展。国外的扩展方法主要有两种:一种是把基于语义树计算的概念间语义相似度作为扩展标准[3],另一种是基于语义模型[4]中概念的共现频率。国内的研究发展相对比较滞后,主要还处于翻译国外研究成果的阶段,其中取得成果较明显的有清华大学计算机科学与技术系智能技术与系统国家重点实验室和哈尔滨大学信息检索室。学者桑艳艳[5]、黄名选[6]都从语义计算的角度分别提出了基于语义关联树的查询扩展算法和一种计算用户查询语义与文档的语义关联权重的新方法,避免了传统的相似度矩阵计算工作量,但在基于概念相似度的扩展过程中都是根据相似度对扩展词加权,所加的权值只能反映概念间的关联程度,却不能反映扩展词与整个查询条件的关联性。学者张敏[7]提出词与词之间的语义关系进行扩展和替换的文档重构方法,实质是一个信息聚集过程,其优点是有目的性扩展且不需要确定查询词和扩展词间相似度的权重,但如果用户查询词刚好是语义树的叶子结点,根据此算法会将其父结点聚集到查询文档中,这样反而扩大查询范围影响检索结果的查准率。文中针对上述的缺陷将相似度和用户查询条件相结合给扩展词加权,采用该加权方式可以修正用户的查询条件,也避免了扩展集合远离用户想表达的查询意欲。

2 语义概念查询扩展基本理论

查询扩展技术发展到现在已有 30 多年的历史,根据计算查询关键词与扩展概念相关度的方法的不同,可以将查询扩展的方法大致分为全局分析法和局部分析法。查询扩展主要的扩展方法有两种:一是根据用户输入的查询关键词,构造与查询关键词意义相近的概念词表。例如用户的查询关键词是“电脑”,可以将“计算机”和“微机”添加到概念的扩展词表中。二是根据概念间关系在实现同义扩展的基础上引进一些推理关系,并实现上下位、平级及蕴涵关系扩展,形成更丰富的、完整的扩展概念集[8]。例如:用户的查询关键词是“电脑”,可以把“计算机”、“微机”、“声卡”、“硬盘”等添加到概念词表中。然而,扩展前需要将概念间的关系量化。目前在信息检索领域用概念关联度来衡量概念间的联系,从自然语言的角度来讲主要考虑语义相关度和语义相似度两方面因素。

定义 1 语义相似度是指概念在意义上的相符合程度,在语义树中通过概念的语义距离计算语义相似度,概念的语义距离与语义相似度成反比。设有概念 C_1 和 C_2,其语义相似度计算方法为:

$$\mathrm{sim}(C_1,C_2)=1-t\sqrt{\frac{1}{2}\alpha\mathrm{Dist}(C_1,C_2)}$$

其中:$\alpha=\dfrac{\mathrm{Dep}(C_2)}{\mathrm{Dep}(C_1)+\mathrm{Dep}(C_2)}$,$\mathrm{Dep}(C)=\sum_{i=1}^{n}1$,

$$\mathrm{Dis}(C_1,C_2)=\sum_{i=1}^{n}\frac{1}{\mathrm{Wid}(C_1)}\frac{1}{2^{\mathrm{Dep}(C_i)}}$$

定义 2 语义相关度是指概念在语义上的关联词程度,如“医生”和“病人”则属于关联关系。语义关联关系在语义树中表示为两个概念间是否存在路径,路径的长度表明概念间关联程度。设有概念 C_1 和 C_2,语义相关度计算法为:

$$\mathrm{Relativity}(C_1,C_2)=\frac{r}{\mathrm{length}(C_1,C_2)+r}$$

其中 r 为调节参数。

定义 3 语义关联度受语义相似度和语义相关度的共同影响,概念 C_1 与 C_2 间的关联度可表示为:

$$\mathrm{Crelevancy}(C_1,C_2)=\alpha\mathrm{Sim}(C_1,C_2)+\beta\mathrm{Relativity}(C_1,C_2)$$

其中 $\alpha+\beta=1$。

3 权值计算方法和算法实现

扩展概念的权值计算是本算法的核心问题,也是文中要阐述的重点。目前较普遍的加权方法是直接将词与词之间的相似度作为权值,能反映出词与词之间的关联程度,但并不能明确地说明扩展词和查询条件间的关系。因此,文中尝试着将扩展概念同用户查询中的所有概念作比较,根据相关度不同设置相应的调节参数值来改变权值大小。具体的计算方法如下:设用户初始的查询条件根据 RDF[9] 三元组形式 (S,P,O) 提取出查询概念集合是 $C=((S_1,S_2,\cdots,S_n),(P_1,P_2,\cdots,P_n),(O_1,O_2,\cdots,O_n))$,其中 S_i、P_i、O_i 分别是三元组中第 i 个概念。根据词与词间的语义关联

度对查询概念集合作语义扩展,将平级词、上下位词以及关联词都扩展进来组成新的概念集合是:

$C' = ((S_{11}^{\wedge W11}, S_{12}^{\wedge W12}, \cdots, S_{1n}^{\wedge W_s 1n}, \cdots, S_{n1}^{\wedge W_s n1}, S_{n2}^{\wedge W_s n1}, \cdots, S_{nn}^{\wedge W_s n1}), (P_{11}^{\wedge W_p 11}, P_{12}^{\wedge W_p 12}, \cdots, P_{1n}^{\wedge W_p 1n}, \cdots, P_{n1}^{\wedge W_p n1}, P_{n2}^{\wedge W_p n2}, \cdots, P_{nn}^{\wedge W_p nn}), (O_{11}^{\wedge W_o 11}, O_{12}^{\wedge W_o 12}, \cdots, O_{1n}^{\wedge W_o 1n}, \cdots, O_{n1}^{\wedge W_o n1}, O_{n2}^{\wedge W_o n2}, \cdots, O_{nn}^{\wedge W_o nn})$

其中 $W_S ij$、$W_P ij$、$W_O ij$ 分别是查询概念集 S、P、O 中第 ij 个元素的权值。

则权值计算表达式:$W(C_{ij}, Q) = 2^K \times \mathrm{AVG}(\mathrm{Crelevancy}(C_{ij}, C_i)) + \bar{\omega}$,$C_{ij}$ 是扩展概念,C_i 是用户查询概念,$\bar{\omega}$ 是扩展阈值,k 取值反映扩展概念与用户查询原意的关联性。由权值计算式可知道参数 k 的取值对权值影响较大,即扩展概念与用户查询意欲的关联性越大,权值相对来说也会越大。为了能得到扩展概念与用户查询意欲的关联性,把扩展概念分别与源扩展词和查询初始条件作布尔运算,根据运算结果确定参数 k 取值。在算法实现时可设置以下三个查询条件:

$Q_1 = (S_i \mathrm{and} S_{ij}) \mathrm{and} (S_i \mathrm{and} P_1 \mathrm{and} P_2 \cdots \mathrm{and} P_n \mathrm{and} O_1 \mathrm{and} O_2 \cdots \mathrm{and} O_n)$

$Q_2 = (S_i \mathrm{and} S_{ij}) \mathrm{and} (S_i \mathrm{or} P_1 \mathrm{or} P_2 \cdots \mathrm{or} P_n \mathrm{or} O_1 \mathrm{or} O_2 \cdots \mathrm{or} O_n)$

$Q_3 = (S_i \mathrm{and} S_{ij}) \mathrm{or} (S_i \mathrm{or} P_1 \mathrm{or} P_2 \cdots \mathrm{or} P_n \mathrm{or} O_1 \mathrm{or} O_2 \cdots \mathrm{or} O_n)$

按照上述设定的查询条件,参数 k 可按下面的方式确定:

$$k = \begin{cases} 0, \text{其他} \\ 1, \text{如果 } Q = Q_3 \\ 2, \text{如果 } Q = Q_2 \\ 3, \text{如果 } Q = Q_1 \end{cases}$$

改进后的查询扩展算法:

对于初始查询条件的三元组形式中每个源关键词 C_i|

用 wordnet 扩展出上下位词、平级词和蕴涵关系词结合 expansion_ set;

对于 expansion_ set 集合中每个元素 C'_i|

计算 C'_i 与 C_i 的语义关联性;

计算 C'_i 与用户初始查询条件的关联性得到参数 k 的值;

根据公式 $W(C_{ij}, Q) = 2^K \times \mathrm{AVG}(\mathrm{Crelevancy}(C_{ij}, C_i)) + \bar{\omega}$ 计算出 k 值;

|

得到所有扩展词的权值集合 weight_ set;

对 weight_ set 中每个元素|

设定权值的阈值删除相关性很小的扩展词,返回符合条件的扩展词;

|

将经筛选后的扩展词及源查询条件提交给检索系统;

|

4　实例分析

从学校图书馆里取 1000 篇文档作为检索源,选取 Wordnet 中局部语义树作为扩展的语义概念空间(见图 1)。假设用户输入查询条件是:Q = 苹果电脑,根据 RDF 三元组形式提取概念得到查询概念集合为 C = (苹果,电脑,苹果电脑)。根据局部语义树得到其扩展集合并计算每个扩展词的关联度,C' = (苹果树 0.5,苹果花 0.5,联想电脑 0.5,苹果显示器 0.5,苹果主板 0.5),如果按照常规的扩展算法将无法筛选出符合条件的扩展集合。按照文中阐述的权值计算方法重新给扩展词加权,得到新的集合是:C' = (苹果树 0.83,苹果花 0.83,联想电脑 1.83,苹果显示器 4.13,苹果主板 4.13),由实验可得当 $W_S ij \geqslant 2$ 时试验结果比较符合实际需要,经过筛选后实际提交给检索系统的查询词有(苹果电脑,苹果显示器,苹果主板)。比较文中的扩展方法与字典法得到的结果如表 1 所示。

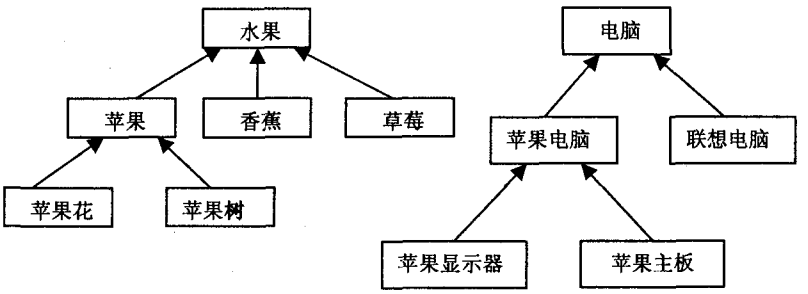

图 1　语义树

表 1　实验结果

算法	前 10 篇中符合条件文档数	前 10 篇查准率	前 100 篇中符合条件文档数	前 100 篇查准率
词典法	7.86	78.6%	76.3	76.3%
文中算法	9.45	94.5%	78.1	78.1%

从实验结果来看,文中算法的查准率在前 100 篇文档中有所降低,其主要原因是可能是实验选取的检索源数量相对较小,影响了整查准率。在今后的研究中将针对此缺陷,选取更大测试集进行实验,进一步验证该算法的实际可行性,力争能更好地提高检索的查准率和查全率。

5　结束语

文中打破了常规的依据语义相似度给扩展词加权

(下转第 7 页)

图 2 显示的是当 PSO 预处理神经网络时,其最大得分情况随训练次数变化的曲线图,相关参数为:训练样本=100,NNG 规模=100, NG 规模=100, 秩代次数=100。从图上可以看出,当训练次数为 60 次的时候,效果最好,这说明当训练样本数量一定的时候,训练次数并不是越多越好,当训练次数过度时,会造成 FNN 对训练样本的过分依赖,减少了种群的多样性,而如果训练次数不足,则达不到最佳的效果。

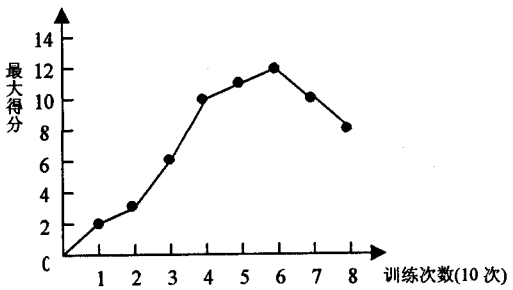

图 2　使用 PSO 预处理方式时,最大得分情况随训练次数变化的曲线图

6　结束语

在机器博弈中,使用 SANE 可以有效地协同进化作为估值函数的神经网络,在这一过程中,并不需要人为干预,只需要一些 AI 程序作为博弈对手即可。实验结果表明:通过 PSO 算法对神经网络进行提前训练后的效果会好得多,但训练时的次数一定要适中,即要必须能在保持种群多样性的同时尽量进行训练学习。下一步的工作包括:尝试更大尺寸的棋盘,改用其他的学习算法对神经网络进行预处理等等。

参考文献:

[1] de Jong D, Pollack J B. Ideal evaluation from coevolution[J]. Evolutionary Computation Journal, 2004, 12(2): 5 - 9.

[2] Axelrod R. The Evolution of Strategies in the Iterated Prisoner's Dilemma[M]//Genetic Algorithms and Simulated Annealing. L. D Ed. [s. l.]: [s. n.], 1987: 32 - 41.

[3] Miller J H. The Coevolution of automata in the repeated prisoner's dilemma[J]. Journal of Economics Behavior and Organization, 1996(29): 87 - 112.

[4] Chellipilla K, Fogel D B. Evolution, Neural Networks, Games, and Intelligence[J]. Proc. IEEE, 1999, 87(9): 1471 - 1498.

[5] Chellipilla K, Fogel D B. Anaconda Defeats Hoyle 6 - 0: A Case Study Competing an Evolved Checkers Program against Commercially Available Software[C]// In Proceedings of Congress on Evolutionary Computation. Piscataway, NJ: IEEE Press, 2000: 857 - 863.

[6] Fogel D B. Blondie24 : Playing at the Edge of AI[M]. Morgan Kaufmann, 2001.

[7] Tripathi A R, Ahmed T, Karnik N M. Experiences and Future Challenges in Mobile Agent Programming[J]. Microprocessors and Microsystems, 2001(25): 121 - 129.

[8] Moriarty D, Miikkulainen R. Discovering complex Othello strategies through evolutionary neural networks[J]. Connection Science, 1995(7): 195 - 209.

[9] Moriarty D, Miikkulainen R. Forming Neural Networks through Efficient and Adaptive Co - Evolution[J]. Evolutionary Computation, 1998, 5(4): 373 - 399.

[10] Moriarty D, Miikkulainen R. Evolving obstacle avoidance behavior in a robot arm[C]//The Fourth International Conference on Simulation of Adaptive Behavior. Cambridge, MA: The MIT Press, 1996: 468 - 475.

(上接第 3 页)

的方式,结合用户初始查询意欲对扩展词加权,使得扩展概念的权值能更直接地反映出此概念与用户初始查询意欲间的关联性。笔者认为,根据此算法计算得到的权值对扩展词进行过滤比直接根据概念词在语义树中的层次和语义距离对概念集合筛选更切合实际。实验表明,此扩展方法已达到预期的效果,基本满足了实际应用要求。

参考文献:

[1] van Rijsbergen. A new theoretical framework for information retrieval[C]// In: Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval. [s. l.]: ACM Press, 1986: 194 - 200.

[2] 张玉叶,李　连, 王春歆. 应用概念语义分析对用户兴趣建模[J]. 计算机与信息技术,2004(9): 123 - 125.

[3] Voorhees E M. Query expansion using lexical - semantic relations[C]// In: Proceedings of the 17th ACM SIGIR Conference on R&D in Information Retrieval. Dublin, Ireland: [s. n.], 1994: 61 - 69.

[4] Gao J F, Nie J Y, Wu G, et al. Dependence Language Model for Information Retrieval[C]// In: Proceedings of the 27th ACM SIGIR Conference on Research and Development in IR. [s. l.]: ACM Perss, 2004: 170 - 177.

[5] 桑艳艳,刘培刚,李　勇. 基于语义计算的查询扩展优化研究[J]. 情报学报,2007(10): 704 - 710.

[6] 黄名选,严小卫,张师超,等. 关联语义的概念查询扩展模型[J]. 情报杂志,2007(8): 92 - 95.

[7] 张　敏,宋睿华,马少平. 基于语义关系查询扩展的文档重构方法[J]. 计算机学报,2004(10): 1395 - 1401.

[8] Qiu Y, Frei H P. Concept Based Query Expansion[C]// In: Proceedings of the 16th ACM SIGIR Conference on Research and Development in IR. [s. l.]: ACM Perss, 1993: 160 - 169.

[9] 张　蕾. 语义 Web 本体语言及 OWL 研究[J]. 成都信息工程学院学报,2007(4): 161 - 165.