基于相似性叙词表的模糊集合模型*)

章 旭 石 进 谢 立
(南京大学计算机科学与技术系 南京 210093)

摘 要 传统的模糊集合模型基于词词关联矩阵来实现模糊检索,词词关联矩阵只考虑语词在文献内部的同时出现。本文提出了一个基于相似性叙词表的模糊集合模型,考虑语词与查询之间的相似性,并将查询扩展包含在此模型中,从而在一定程度上提高了检索性能。

关键词 模糊集合模型,相似性叙词表,查询扩展

Fuzzy Set Model Based on Similarity Thesaurus

ZHANG Xu SHI Jin XIE Li
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

Abstract Traditional fuzzy set model is based on term-term correlation matrix for fuzzy retrieval, which only considers term co-occurrence within documents. The paper presents a fuzzy set model based on similarity thesaurus which considers the similarity between term and query. And we add the query expansion into the model. To some extent, the retrieval performance is improved.

Keywords Fuzzy set model, Similarity thesaurus, Query expansion

1 引言

随着信息社会的发展和计算机应用的普及,信息检索已经成为人们日常生活不可缺少的一部分。然而,现有的基于普通集合理论和布尔逻辑的搜索引擎往往不能满足人们的需要。这是因为在现实世界中存在着大量的模糊性信息,例如"电脑"和在学术场合称作的"计算机"虽然表达的是同一个概念,但是在以"电脑"作为关键词进行搜索返回的结果中,或者没有出现包含"计算机"这一关键词的相关文档,或者排序比较靠后。正是为了模拟人脑思维的模糊性特点,创立了模糊集合理论,从而开创了模糊信息处理的新领域。将模糊方法和叙词表技术引入信息检索,可以满足信息社会中人们对信息检索的大量需求。但是,基于词词关联矩阵的模糊集合模型[1]由于存在自身的不足,即只利用词词之间的相关性来进行模糊检索,而没有考虑词与整个查询概念的相关性。本文首先介绍了叙词表的特点及自动构建,接着在传统模糊集合模型的基础上提出了基于相似性叙词表的模糊集合模型,并讨论了如何进行用户查询扩展以改善检索性能,最后对这两种模型进行了比较。

2 叙词表的特点及自动构建

2.1 叙词表的特点

叙词表是一种由概括一门或各个学科领域的、经规范化处理的语义相关和族性相关的词或词组,按特定顺序排列所形成的后组式检索词典。它是一种将标引人员和检索人员的自然语言转换成规范化检索语言的术语控制工具。根据Foskett的观点[2],叙词表的基本目标是:(1)为标引和检索提供标准化的词汇表或参照系统;(2)帮助用户确定哪些词语适合于查询表达式;(3)根据用户需要,提供当前查询上位类和下位类的分类层次。叙词表中语词之间的关系给我们提供了有价值的信息,很自然地,我们想到用叙词表来帮助用户扩展初始查询[3]。

2.2 相似性叙词表的自动构建

由于手工构建的叙词表工作量较大,我们这里采用自动构建的相似性叙词表,并给出相应的构建方法。自动构建的叙词表通常依赖于所用的文本数据库。典型地,自动构建的叙词表是基于共同出现的信息,而且使用相关性判断来估计语词与查询语词或特定查询相似的概率[4],从而扩展查询。

相似性叙词表是由语词之间相似性组成的矩阵,语词扮演了文献的角色,而文献作为语词的索引。这样,我们通过任意的互换文献和语词角色的检索方法来自动构建相似性叙词表[5]。

对每个语词 t_i,用文献向量空间(DVS)中的一个向量 $\vec{t_i}=(w_{i,1},w_{i,2},\cdots,w_{i,n})$ 表示,$w_{i,j}$ 表示文献 d_j 关于语词 t_i 的特征权值,n 是集合中的文献数。$w_{i,j}$ 由如下公式计算得出:

$$w_{i,j}=\frac{(0.5+0.5\frac{f_{i,j}}{\max_j(f_{i,j})})itf_j}{\sqrt{\sum_{i=1}^{n}(0.5+0.5\frac{f_{i,j}}{\max_j(f_{i,j})})^2itf_j^2}} \quad (1)$$

其中,$f_{i,j}$ 表示语词 t_i 在文献 d_j 中的出现频率,itf_j 表示文献 d_j 的逆词频,$\max_j(f_{i,j})$ 表示语词 t_i 在集合所有文献中 $f_{i,j}$ 的最大值。逆词频 itf_j 的计算由以下公式给出:

$$itf_j=\log\frac{N}{N_j} \quad (2)$$

其中,N 表示整个文献集合中语词的数目,N_j 表示文献 d_j 中不同标引词的数目。逆词频的定义表明短文献比长文献起更加重要的作用。如果两个语词共同出现在一个长文献中,那

*)本文受江苏省科学技术厅项目(BA2005001)资助。章 旭 硕士研究生,主要研究方向为操作系统、信息检索;石 进 博士研究生,主要研究领域为操作系统、信息安全等;谢 立 教授,博士生导师,主要研究领域为分布式计算、并行处理、操作系统。

么这两个语词相似的概率要小于它们共同出现在一个短文献中相似的概率。

从式(1)可以推导得出$|\vec{t_i}|=\sqrt{\sum_{j=1}^{n}w_{i,j}^2}=1$，这意味着$\vec{t_i}$是一个单位向量。

有了以上这些定义，两个语词 t_i 和 t_j 之间的相似性度量可以用对应向量之间的点积给出：

$$s(t_i,t_j)=\vec{t_i}\cdot\vec{t_j}=\sum_{k=1}^{n}w_{i,k}\times w_{j,k} \tag{3}$$

相似性叙词表通过计算出所有语词对之间的相似性度量来构建。可以看出，相似性叙词表是一个对称矩阵，元素在如下范围内取值：$0\leqslant s(t_i,t_j)\leqslant 1$。对于大型数据库来说，构建这样一个相似性叙词表的计算量是巨大的。然而，只需要计算一次，因为给一个大型数据库增加少量的文献几乎不会改变语词之间的关系。而且，通过修改对应于包含在新文献中语词的入口可以获得相似性叙词表的更新。更精确地说就是重新估计新到来的语词之间的相似性，然后更新相似性叙词表相应的入口。

3 模糊集合模型

模糊集合模型已经被用来定义模糊查询，即查询语词和文献之间的模糊关系。每一个查询语词定义了一个模糊集合，每个文献在相应的语词集合中有一个隶属度，取值通常小于 1。继续讨论之前，首先介绍一些基本概念[1]。

3.1 模糊集合理论

模糊集合理论[6]研究的是边界不明确的集合表示，其中心思想是把隶属函数和集中的元素结合在一起。该函数的取值在区间[0,1]上，0 对应于不隶属于该集合，1 表示完全隶属于该集合，隶属值在 0 和 1 之间表示集合中的边际元素。

定义 1 设在论域 U 上给定映射 $\mu:U\to[0,1]$，则说 μ 确定了 U 上的一个模糊子集，记为 A。μ 称为 A 的隶属函数，记为 $\mu_A(u)$。

对 $\forall u_0\in U$，$\mu_A(u_0)$ 称为元素 u_0 关于 A 的隶属度，它表示元素 u_0 隶属于 A 的程度。

定义 2 设 U 表示论域，A 和 B 分别表示 U 的两个模糊子集，$\overline{A}$是 A 关于 U 的补集，μ 表示 U 的元素，则

$$\mu_{\overline{A}}(u)=1-\mu_A(u)$$

$$\mu_{A\cup B}(u)=\max(\mu_A(u),\mu_B(u))$$

$$\mu_{A\cap B}(u)=\min(\mu_A(u),\mu_B(u))$$

3.2 基于相似性叙词表的模糊集合模型

基于传统模糊集合模型的思想，我们以相似性叙词表代替词词关联矩阵来构建模糊集合模型，并考虑查询扩展以查询检索出额外的相关文献。我们也用代数的方法来计算文献属于符合查询的模糊集合的隶属度。

类似地，在此模型中与每个标引词 t_i 相关联的模糊集合可以由以上构建的相似性叙词表来定义，在这个集合中，文献 d_j 的隶属度可以计算如下：

$$\mu_{i,j}=1-\prod_{t_i\in d_j}(1-s_{i,j}) \tag{4}$$

如果文献 d_j 自身的语词与 t_i 有关，则该文献属于语词 t_i 的模糊集合。只要文献 d_j 中至少有一个标引词 t_l 与标引词 t_i 密切相关(如 $s_{i,j}\approx 1$)，则 $\mu_{i,j}\approx 1$；如果文献 d_j 中所有的标引词与 t_i 不是密切相关的，则 $\mu_{i,j}\approx 0$。

基于相似性叙词表的模糊集合模型把表示用户信息需求的布尔查询表达式转换成析取范式。例如，考虑查询$[q=t_a\wedge(t_b\vee\neg t_c)]$，可以写成析取范式的形式：$[q_{dnf}=(1,1,1)\vee(1,1,0)\vee(1,0,0)]$，其中每一个分量都是三元组$(t_a,t_b,t_c)$的一个二值加权向量。用 ss_i 表示第 i 个分量，则 $q_{dnf}=ss_1\vee ss_2\vee ss_3$。用 D_a 表示与索引 t_a 相关联的文献模糊集合，比如该集合由隶属度 $\mu_{a,j}$ 大于给定阈值 T 的文献 d_j 组成；用$\overline{D_a}$表示 D_a 的补集，表示与$\overline{t_a}$相关联的模糊集合。类似地，可以分别定义标引词 t_b 的模糊集合 D_b 和标引词 t_c 的模糊集合 D_c。

查询模糊集合 D_q 是与 q_{dnf} 三个分量相关联的模糊集合的并集，即 $D_q=(D_a\cap D_b\cap D_c)\cup(D_a\cap D_b\cap\overline{D_c})\cup(D_a\cap\overline{D_b}\cap\overline{D_c})$。模糊结果集合 D_q 中文献 d_j 的隶属度可以通过以下公式来计算：

$$\begin{aligned}\mu_{q,j}=\mu_{ss_1+ss_2+ss_3,j}&=1-\prod_{i=1}^{3}(1-\mu_{ss_i,j})=\\&1-(1-\mu_{a,j}\mu_{b,j}\mu_{c,j})\times(1-\mu_{a,j}\mu_{b,j}(1-\mu_{c,j}))\times\\&(1-\mu_{a,j}(1-\mu_{b,j})(1-\mu_{c,j}))\end{aligned} \tag{5}$$

此模糊集合模型使用相似性叙词表计算文献 d_j 与模糊标引词之间的相关性，此外还计算了由用户查询定义的模糊集合中文献 d_j 的全部隶属度以进行查询结果的排序。然而其不足主要体现在只考虑语词之间的相似性，没有看到语词与查询之间的相似性，从而不可避免会出现与整个查询密切相关的文献未被检索到。为了解决此问题，我们必须通过额外的方法进行查询语词的扩展，这些语词必须满足的条件就是与查询的相似度要大于某个阈值 K。下面讨论如何基于相似性叙词表选择相关语词扩展查询以完善上述模糊集合模型。

查询 q 用语词向量空间中的向量 $\vec{q}=(w_{1,q},w_{2,q},\cdots,w_{n,q})$来表示，语词向量空间由结合中的所有语词定义。这里，$w_{i,q}$是查询 q 关于语词 t_i 的权值，该权值的计算类似于式(1)。既然相似性叙词表表示的是文献向量空间的语词之间的相似度，就需要把语词向量空间的向量 $\vec{q}$ 映射到文献向量空间中。

定义 查询 q 与文献向量空间中的向量$\vec{q_c}$相互关联，则 $\vec{q_c}=\sum_{t_i\in q}w_{i,q}\vec{t_i}$。

从而语词 t 与查询 q 之间的相似度 $s(q,t)$计算如下：

$$s(q,t)=\vec{q_c}\cdot\vec{t}=(\sum_{t_i\in q}w_{i,q}\vec{t_i})\cdot\vec{t}=\sum_{t_i\in q}w_{i,q}\times s(t_i,t) \tag{6}$$

根据预先设定的阈值 K，将与查询 q 相似度大于此阈值的语词加入到初始查询中。假设有 m 个语词 $k_1,k_2,\cdots,k_m$ 加入到初始查询中，则将这 m 个语词所对应的模糊集合作为结果返回给用户。结果文献的隶属度可以通过以下式子求得：

文献 d_j 在语词 k_i 相关联的模糊集合中的隶属度为

$$\mu_{i,j}=(1-\prod_{t_l\in d_j}(1-s_{i,l}))\times s(q,k_i) \tag{7}$$

利用此隶属度可以对检索结果进行排序。

传统的模糊集合模型只考虑利用语词共现关系来进行模糊检索，并给出了文献隶属度的计算。由于它没有考虑语词与整个查询概念的相似性，模糊程度有限。采用相似性叙词表的模糊集合模型能够充分发挥叙词表的优点[5]：

1)为大集合创建的相似性叙词表的质量比为小集合创建的叙词表的质量要高，随着集合内文献数量的增多，检索效率也会随着增加。

2)依赖相关反馈信息的方法仅仅在检索出的一部分文献中选择查询扩展所需的语词，相比较而言，基于相似性叙词表

(下转第 229 页)

利用交互过程的动态信息，采用基于对方议题保留值估计值的提议策略，Agent a_1 与 Agent a_2 协商的结果如表 3。

表 3　未学习历史信息的协商结果

协商	a_1(卖方) $V^z_{m\,a_1}$ $\bar{V}^z_{m\,a_1}$	a_2(买方) $V^z_{m\,a_2}$ $\bar{V}^z_{m\,a_2}$	协商次数	达成一致协议值
S_1	⟨120,57⟩ ⟨80,45⟩	⟨72,40⟩ ⟨90,50⟩	4	⟨80,46⟩
S_2	⟨120,57⟩ ⟨90,40⟩	⟨72,40⟩ ⟨95,43⟩	9	⟨92,40⟩
S_3	⟨120,57⟩ ⟨80,40⟩	⟨72,40⟩ ⟨100,55⟩	5	⟨82,40⟩
S_4	⟨120,57⟩ ⟨90,41⟩	⟨72,40⟩ ⟨92,45⟩	9	⟨92,40⟩
S_5	⟨120,57⟩ ⟨100,45⟩	⟨72,40⟩ ⟨95,50⟩	11	⟨95,49⟩

以 Agent 成功的交互历史信息和交互过程中的动态信息为依据，采用基于对方议题保留值估计值的提议策略，Agent a_1 与 Agent a_2 协商的结果如表 4。

表 4　学习历史信息的协商结果

协商	a_1(卖方) $V^z_{m\,a_1}$ $\bar{V}^z_{m\,a_1}$	a_2(买方) $V^z_{m\,a_2}$ $\bar{V}^z_{m\,a_2}$	协商次数	达成一致协议值
S_1	⟨120,57⟩ ⟨80,45⟩	⟨72,40⟩ ⟨90,50⟩	2	⟨82,45⟩
S_2	⟨120,57⟩ ⟨90,40⟩	⟨72,40⟩ ⟨95,43⟩	5	⟨90,41⟩
S_3	⟨120,57⟩ ⟨80,40⟩	⟨72,40⟩ ⟨100,55⟩	2	⟨83,40⟩
S_4	⟨120,57⟩ ⟨90,41⟩	⟨72,40⟩ ⟨92,45⟩	6	⟨90,42⟩
S_5	⟨120,57⟩ ⟨100,45⟩	⟨72,40⟩ ⟨95,50⟩	8	⟨94,48⟩

利用表 3、表 4 中的协商结果计算 Agent a_1 与 Agent a_2 的联合效用，即买卖双方多议题整体效用之和进行比较，如图 1 所示。横坐标表示协商，纵坐标表示协商的相对的联合效用。

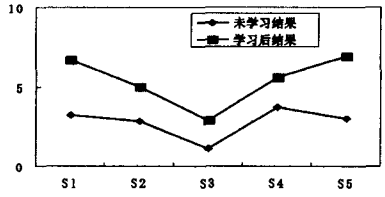

图 1　学习历史信息前后效用比较

以上实验结果表明，利用成功的协商历史信息和交互过程中的动态信息可以较准确地把握协商对手的初始信念，提高协商双方的整体效用，加快协商进程。

结束语　本文希望能从 Agent 协商的中间平台上获取交互历史信息，然后通过分析 Agent 个体的成功协商历史来提取出使协商项目效用最大化的信息作为协商双方 Agent 的初始信念，从而在协商开始前较为准确预测对方 Agent 的保留值。在协商过程中，查找出差异较大的协商议题，根据交互历史信息与交互过程中动态信息，利用贝叶斯公式，通过在线学习机制预测对手信念并调整己方策略。在此基础之上提出了协商框架，优化协商进程，探讨希望能最大化双方的收益。由于多 Agent 系统所处环境的复杂性，Agent 个体的偏好、友好度及议题间是否均存在可补偿问题将是进一步讨论的问题。

参 考 文 献

[1] Wooldridge M. Agent-based Software Engineering. IEEE Proceedingson Software Engineering, 1997, 144(1): 26-37

[2] Rahwan I, Ramchurn S D, Jennings N R, et al. Argumentation-based negotiation. Knowledge Engineering Review, 2004: 343-375

[3] Jennings N R, Parsons S, Sierra C, et al. Automated Negotiation // Proceedings of the 5th International Conference on the Practical Application of Intelligent Agents and Multi-Agent Systems (PAAM-2000). 2000: 23-30

[4] Matos N, Sierra C, Jennings N R. Determining successful negotiation trategies: an evolutioanary approach // Proceedings of the 3rd International Conference on Multi-Agent Systems (ICMAS'98). 1998: 182-189

[5] Zeng D, Sycara K. Bayesian Learning in negotiation. International Journal of Human-Computer Studies, 1998, 48: 125-141

[6] Faratin P, Sierra C, Jennings N R. Negotiation decision functions forautonomous Agents. International Journal of Robotics and Autonomous Systems, 1998: 24(3/4): 5-19

[7] Uprea MElectronic An Adaptive Negotiation Model for Agent-Based Commerce. DJ. Romania: Department of Informatias. University of Ploiesti, 2000

[8] 李勇，李石君. 多 Agent 自动协商. 计算机工程，2003，29(6)：59-63

[9] 王黎明，黄厚宽. 一个基于多阶段的多 Agent 多问题协商框架. 计算机研究与发展，2005：1849-1855

[10] 王立春，陈世福. 多 Agent 多问题协商模型. 软件学报，2002，13(8)：1637-1643

[11] 陈亚楠，王黎明. 基于贝叶斯的多议题协商优化. 计算机工程与应用，2006(6)：69-71

[12] 王娟，柴玉梅. 基于多议题协商的贝叶斯学习. 计算机技术与发展，2006，16(2)

[13] 彭志平，彭文，等. 一种双边多议题自治协商模型的研究. 电子与信息学报，2007，29(3)

(上接第 202 页)

的方法能充分利用向量模型的特征，把与初始查询相关度大于某个阈值的语词添加进来，选择语词的范围和数量相对较大，从而能提高文献的查全率。

结束语　针对词词关联矩阵只考虑语词之间共同出现的缺陷，本文提出的基于相似性叙词表的模糊集合模型，通过在相似性叙词表基础上建立模型，并考虑语词与整个查询之间的相关性从而扩展查询，在一定程度上提高了信息检索的性能。在未来的工作中，我们将深入研究叙词表在 Web 上的应用，并对此模型进行进一步的改进。另外，由于商业数据库文献和语词数量庞大，构建一个相似性叙词表开销也是相当大的，因此叙词表的构建算法，以及存储和访问方式也有待进一步研究。

参 考 文 献

[1] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Beijing: China Machine Press, 2004

[2] Foskett D J. Reading in Information Retrieval. Thesaurus. Jones K S, Willet P, eds. Morgan Kaufmannn Publishers Inc., 1997: 111-134

[3] Jing Y, Croft W B. An association thesaurus for information retrieval[A] // Processing of the Intelligent Multimedia Information Retrieval Systems (RIAO'94)[C]. 1994: 146-160

[4] Xu Jinxi, Croft W B. Query expansion using local and global document analysis // Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland, 1996: 4-11

[5] Qiu Y, Frei H P. Concept based query expansion[A] // Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'93)[C]. 1993: 160-169

[6] Zadeh L A. Fuzzy sets. Readings in Fuzzy Sets for Intelligent Systems. Dubois D, Prade H, Yager R R, eds. Morgan Kaufmann, 1993