第16卷 第2期
2006年2月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.2
Feb. 2006

搜索引擎中语义相关反馈技术的研究

殷亚玲,张 蕾
(西北大学 计算机科学系,陕西 西安 710069)

摘 要:搜索引擎是互联网普及的标志,但是目前在搜索引擎的召回率和准确率上是不能让用户满意的。文中从基于向量空间的反馈技术的分析入手,利用概念图的知识,提出了基于语义的相关反馈技术。实验结果表明,该技术在提高查全率和查准率上是有效的。

关键词:搜索引擎;相关反馈;向量空间;概念图
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1005-3751(2006)02-0167-04

Research of Relevance Feedback Based on Semantic in Search Engine

YIN Ya-ling, ZHANG Lei
(Department of Computer Science, Northwest University, Xi'an 710069, China)

Abstract: Search engine is the sign of World Wide Web prevalence. But now the precision and the recall is not satisfied for the users. Firstly the paper analyzes the defect of technique of the VSM model. And then put forward one technique based-on semantic for relevance feedback, using the knowledge of conceptual graphs. Experimental results show that the technique can help user accomplish the task effectively.

Key words: search engine; relevance feedback; VSM; conceptual graphs

0 引 言

目前常用的搜索引擎技术是基于关键词匹配的技术,但是由于缺乏语义信息,检索手段单一,存在检索要求难以表达,返回结果不相关,质量低,返回结果展现方式单调,在查全率和查准率上精度不够等问题。比如查找计算机,电脑也是计算机的信息就不能被检索出来。又如查找土豆,土豆属于蔬菜的信息不能被挖掘出来。用如何使深蓝更蓝在百度中检索,前十条信息没有一条是关于计算机发展方向的。目前解决上述问题的方法主要是在搜索引擎中加入相关反馈的方式和建立语义 Web 的方式。文中采用在搜索引擎中加入含有语义信息的相关反馈技术来提高检索的效率。

相关反馈技术(relevance feedback)是检索模型的重要组成部分,目的是将用户查询到的信息反馈给系统,系统根据这些信息和用户的需求对提问式进行调整,使调整后的信息更好地近似于用户需要的信息的这样一个过程。目前用在搜索引擎中的反馈技术主要有基于向量空间模型的反馈技术和基于统计的反馈技术,主要采用关键词匹配的方式,在反馈过程中通过用户的判断决定检索出的文档的相关性,根据一定的权值算法扩展提问式,这在提高查全率和查准率上是有效的,但是不能从根本上解决概念深层的语义问题。

收稿日期:2005-05-18
基金项目:陕西省教育厅专项科研基金(HD01302)
作者简介:殷亚玲(1979—),女,陕西杨凌人,硕士研究生,研究方向为人工智能及自然语言理解;张 蕾,教授,研究方向为人工智能及自然语言理解。

1 背 景

1.1 概念图

概念图[1,2]是一种描述复杂对象结构的知识表示工具,其思想来源于 C.S.Pierce 的存在图和菲尔墨的语义网络,其理论建立在谓词逻辑上,能完全与自然语言相互翻译,表示出自然语言的语义。概念图一般由概念和关系组成,表示方式有 Linear notation 和 graphical notation 两种。概念图是一个二分图,表示概念和关系之间用弧连接,而概念和概念之间,关系和关系之间没有弧连接,弧属于关系但是依附于概念。

定义1 概念:[Type:Referent][3]

其中 Type(概念类型),Referent 抽取由具体的文本分类算法决定,一般为实词。Referent 可以为空,Type 不能为空。比如概念[Person:John]中 Person 指的是 Type 而 John 指的是 Referent,概念[Bus]中的 Referent 为空。Type 可以被分为 super type(父类)和 sub type(子类),Referent 定义为 Type 的实例。

定义2 关系:关系由3部分组成,即 relation type、valance 和 signature[3]。

relation type 根据用户需要定义，常用的有 Agnt(Agent)，Thme(Theme)等 23 种。

Valance 表示的是关系所连接的弧的数量，一般大于或等于 1；依据 valance 的数量将关系成为一元关系、二元关系和 n 元关系。

signature：n 元关系 r 的 signature 表示成 $< t_1, t_2, \cdots, t_n >$，其中 t_1, t_2, t_3 等分别为属于关系 r，依附于概念的弧的概念类型 Type。

比如因为有[Sing]→(Agnt)→[Bird]，而 Sing 和 Bird 分别为 Act 和 Animal 的子类，故此关系 Agnt 的 signature 为 < Act, Animal >。

1.2 向量空间模型中的相关反馈技术

参照文献[4]可以看到一般的反馈模型主要还是通过从用户那里接受相关性评估，输出相关文档和不相关文档，实现相关反馈公式，目的是修改提问式。在向量空间模型中提问式的优化和修改的基本思路为：

(1)提问式修改的基本思路。

出现在相关文档中的词项或权值增长的词项被添加到原始的提问式向量中；出现在不相关文档中的或权值减轻的词项被从原始提问式中删除。理想的情况是提问式的词项只出现在相关文档中。但是一个词项可能在相关文档和不相关文档中同时出现，那么在提问式中是否要包含该词项，一般采用提问式优化的思路解决。

(2)提问式优化的思路。

首先将文档划分成相关和不相关两个集合，分别定义词项在两个集合中的权值，然后将它们添加到提问式修改公式中可以解决在提问式中是否要包含该词项的问题，但是相关文档和不相关文档的划分却无法解决。

基于这种情况，文中提出基于语义的反馈技术，将文档表示成为概念图的形式，在图中通过概念之间的距离计算概念之间的相似度，进行提问式扩展，避免了上述问题。

2 语义相关反馈模型

2.1 概念图的改进

文中在文献的基础上结合概念分类学与原始概念图的定义对概念图进行改进，目的是在清楚的层次结构中有效地描述概念，简化概念之间相似度的计算，将其从逻辑推理的角度转化为数字计算的方式。

定义 3 概念：由文档或信息的实词构成，分为原子概念(atom concept)和组合概念(compound concept)，每一个组合概念都是由两个或两个以上原子概念组合而成，也可由原子概念和组合概念结合形成新的组合概念。概念的定义同定义 1 中的 Type 和 Referent，将 Type 表示成 Referent 的父概念，通称为 Type。

例如 black、dog 两个原子概念表示成组合概念 dog[CHR:black]

定义 4 关系：由 relation type、valance 和 signature 三部分组成，定义同定义 2。关系分为两类：ISA 关系(concept inclusion)和其它语义关系，其中概念和概念之间的 ISA 关系在图中可以省略。其它语义关系包含有 WAT(With - respect - to)，CHR(Characterized - by) CBY(Caused - by) TMP(Temporal)等。

设 R 表示关系的集合，即 $R = \{$WAT, CHR, CBY, TMP, LOC, $\cdots\}$，$r_i \in R, y_i \in R, i = 1, \cdots, n$。

定义 5 对于两个原子概念 x, y，如果 x ISA y，表示为 $x \leqslant y$，那么 x, y 扩展的组合概念有 $x[\cdots] \leqslant y[\cdots]$、$x[\cdots, r:z] \leqslant y[\cdots]$ 和 $x[\cdots, r:z] \leqslant y[\cdots, r:z]$，其中 $r \in R, z$ 为概念。

定义 6 如果原子概念 x, y 有 $x \leqslant y$，那么概念 z 扩展后就有 $z[\cdots, r:x] \leqslant z[\cdots, r:y]$，

如果 $x[r_1:y_1, \cdots, r_n:y_n] \leqslant y$，那么 $x[r_1:y_1, \cdots, r_n:y_n, r_{n+1}:y_{n+1}] \leqslant y$

定义 7 如果 x ISA y，那么概念 x 到概念 y 称为泛化(Generalization)，用 g 表示，概念 y 到概念 x 成为特化(Specialization)，用 s 表示。

显然，泛化相对特化要容易些，图 1 中 dog ISA animal，dog 指向 animal 是泛化，是正确的，而 animal 指向 dog 是特化，就不一定是正确的。基于上面描述例子：the black dog is making noise 就表示成概念图：Noise [CBY:dog[CHR:black]]。图 1 是改进后定义的概念图的一部分，其中 ISA 关系被省略，ρ_{CHR} 和 ρ_{CBY} 分别表示的是 CHR 和 CBY 关系。

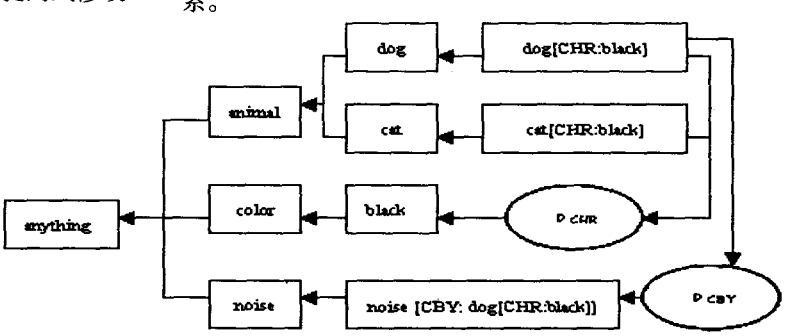

图 1 概念图的一部分

2.2 相似度的计算方式

根据文献[5]，概念图的匹配方式有最大连接法、限制合一法等，这些方法都是运用逻辑推理的方式，文中在改进的概念图上采用数字计算的方式进行相似度的计算[6]。首先进行如下假设和定义：

定义 8 $\tau(c)$ 表示的是概念 c 的扩展，$\omega(c)$ 表示的是概念 c 向上扩展 $c = c_0[r_1:c_1, \cdots, r_n, c_n]$，subterm$(c) = \{c_0, c_1, \cdots, c_n\}$，将 subterm 定义为一系列概念的集合，即 subterm$(c) = \bigcup_i$subterm(c_i)，那么扩展的概念 $\tau(c) = \{c\} \cup \{x \mid x \in \text{subterm}^k(c)\}$，即 $\tau(c) = \{c \leqslant x \vee c \leqslant y[r:x], x \in L, y \in L, r \in R\}$

定义 9 $\omega(c) = \{x \mid x \in c \vee y \in c, y \text{ISA} x\}$，

$\gamma(c) = \{(x,y,\mathrm{ISA}) \mid x,y \in \omega(\tau(c)), x\mathrm{ISA}y\} \cup \{(x,y,r) \mid x,y \in \omega(\tau(c)), r \in R, x[r:y] \in \tau(c)\}$

那么相似度函数为:

$$\mathrm{sim}(x,y) = \rho \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|} + (1-\rho) \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(y)|} \quad (1)$$

其中:$\alpha(c) = \in \omega(\tau(c)), \rho \in [0,1]$ 表示泛化的程度。

那么 τ(noise[CBY:dog[CHR:black]]) = {noise[CBY:dog[CHR:black]], noise[CBY:dog], noise, dog[CHR:black], dog, black},显然 sim(cat[CHR: black, CHR: brown], dog[CHR: black, CHR:brown]) > sim(cat, dog)

定义 10　设 T 为概念,那么以 T 为提问式,就可以被扩展为:

$T += 1/T + \mathrm{sim}(T,T_1)/T_1 + \cdots + \mathrm{sim}(T,T_n)/T_n$,其中 $T_1,\cdots,T_n$ 为与 T 具有相似性的概念。

如果对于两个概念的扩充有多个路径那么就采用定义 11。

定义 11　有 k 个不同的关系,$R^1,\cdots,R^k$, $\rho^1,\cdots,\rho^k$ 为对应的相似参数,$p = (p_1,\cdots,p_n)$ 为路径,$r^j(p)$ 为对应的路径 p 所连的关系 R^i 的数量,那么 $r^j(p) = |\{i \mid p_i R^j p_{i+1}\}|$,如果 $p^1,\cdots,p^m$ 表示两个概念 x,y 之间的所有路径:

$$\mathrm{sim}(x,y) = \max_{j=1,\cdots m} \{\sigma^{s(p^j)} \gamma^{g(p^j)} \rho_1^{r^1(p^j)} \cdots \rho_k^{r^k(p^j)}\} \quad (2)$$

图 1 中:如果 $x\mathrm{ISA}y$,那么 $\mathrm{sim}(x,y) = \gamma, \mathrm{sim}(y,x) = \delta$,根据文献[6]的规定取 $\delta = 0.9$ 和 $\gamma = 0.4$,$\rho_{CHR} = 0.3$,$\rho_{CBY} = 0.2$。那么 sim(cat, dog) = sim(cat, animal) * sim(animal, dog) = 0.4 * 0.9 = 0.36; dog += 1/dog + γ/animal + ρ_{CHR}/dog[CHR:black]。

具有不同路经的两个概念取它们之间相似度最大值:sim(cat[CHR:black], dog[CHR:black]) = max(γ * γ * δ * δ, ρ_{CHR} * ρ_{CHR}) = 0.9

2.3　语义相关反馈技术的模型

2.3.1　思想

(1)用户在搜索引擎中依据提问式进行检索,将提问式通过用户接口表示成为概念图;

(2)对检索出的信息采用聚类的方式进行文本分类、文本切分,抽取其中的关键信息建立概念图,进而形成概念图库;

(3)利用搜索器采用分布式方式进行搜索,将搜索的信息填充到概念图库中,对概念图库进行整理;

(4)将提问式和检索出来的信息进行比较,计算语义相似度,扩充提问式;

(5)在概念图库中对提问式作进一步的语义扩充;

(6)将提问式提交给用户接口,进行进一步检索,重复步骤(2),直到用户满意。

2.3.2　模型

语义检索模型如图 2 所示。

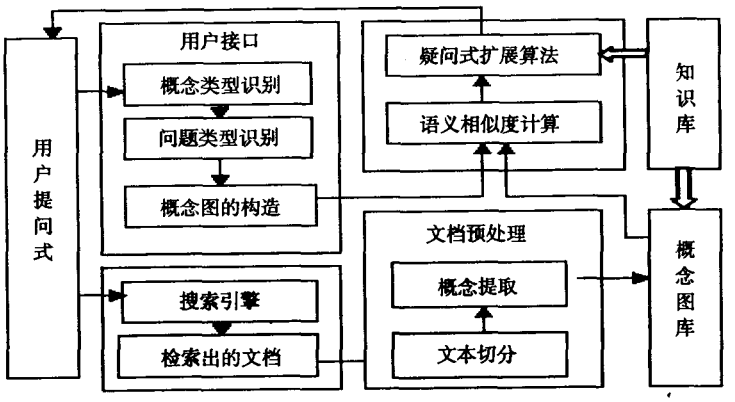

图 2　语义检索模型

相关反馈模型分为 4 部分:

(1)用户接口模块:传统的提问式都是一个或多个关键词,在本模型中允许用户用自然语言作为提问式。比如对于"如何使深蓝更蓝"这样的提问式,用户接口就需要对其进行分析,解析出用户的真正意图是查询"计算机的未来与发展方向"。对于提问式进行分词和词性标注,使用抽词工具结合知识库,识别专有人名、机构名、专业术语等。然后对标注后的词进行句法分析。概念类型识别的作用是根据句法分析结果结合知识库识别出提问式所描述的概念类型的领域。比如:"深蓝"是"计算机"代称,属性"蓝"表示"发展方向"。问题类型的识别是指将用户的提问式根据类型库划分到一个指定的类型中[7,8],便于语义信息的找寻。

(2)检索文档处理模块:主要包含文档切分和概念的提取,参照文献[9]和文献[2]。

(3)概念图库模块:将提取的概念建成概念图库,利用搜索器扩充概念库,在知识库中进行概念图库的修改和管理。

(4)语义相似度计算和提问式扩展模块:利用概念图中的定义,根据公式(1)、(2)计算概念之间的相似度,根据定义 10 扩展提问式。

2.3.3　实验

笔者采用表 1 的提问式在百度和 Google 中进行检索,然后对检索出的文档进行分析。

表 1　提问式及其包含的语义信息

满意度	提问式	查询意图
好	如何使深蓝更蓝	计算机的未来与发展方向
好	师范类自荐书	师范类学生找工作时需要的自荐书
好	文化大革命爆发的原因	了解文化大革命的爆发原因
好	局域网的发展历史及现状	局域网产生的过程及背景
差	米酒的制作方法	怎么制作糯米酒
好	成本领先战略在饭店业应用范围	什么样的饭店适合用成本领先战略
好	幼儿安全教育的研究论文	幼儿自我保护意识和能力的培养

采用准确率和召回率对反馈技术在检索模型上的有效性进行分析，设采用反馈技术得到的与查询相关的一组文档记为{Relevant}，由系统检索出的一组文档记为{Retrieved}，既相关又被检索出的一组文档记为{Relevant}∩{Retrieved}。召回率(Recall)是与查询相关的，并且被实际检索出的文档的百分比。准确率(Precision)是所检索到的实际文档与查询相关的文档的百分比。它们的形式定义分别如下：

$$\text{Recall}=\frac{|\{\text{relevant}\}\cap\{\text{retrieved}\}|}{|\{\text{relevant}\}|}$$

$$\text{Precision}=\frac{|\{\text{relevant}\}\cap\{\text{retrieved}\}|}{|\{\text{retrieved}\}|}$$

准确率反映了系统检索相关文档的专一性，而召回率反映了系统检索所有相关文档的完备性，所以实验中将二者结合起来作为实验结果的评估方式。根据实验结果可以看出基于概念图的反馈技术在信息检索模型上是有效的，但是实验结果同时表明该方法在实践中还需要进一步改进，以期提高准确率和召回率。实验结果见表 2。

表 2 几种反馈技术的实验结果

反馈技术	准确率	召回率
基于向量空间的反馈技术	73.12%	74.25%
基于语义的反馈技术	75.83%	76.12%

3 结束语

文中提出的基于语义的相关反馈技术是反馈技术在语义上一次尝试。用概念图的方法体现了概念之间的语义信息及其关系，将它们用在反馈技术中，是目前反馈技术发展的趋势。在搜索引擎下的实验表明，该方法在提高检索效率，满足用户需求上是有效的。然而反馈技术是一个动态的过程，要想比较有效地利用语义信息，还需要从用户的角度结合人机交互方式进行分析，这也是今后研究的方向。

参考文献：

[1] Sowa J F. Conceptual Structures: Information Processing in Mind and Machine[M]. Reading, MA: Addison - Wesley, 1984.

[2] Sowa J F. Knowledge Representation: Logical, Philosophical, and Computational Foundations [M]. Pacific Grove, CA: Brooks Cole Publishing Co, 2000.

[3] Petersen U. Conceptual Structure[EB/OL]. http://www.huminf.aau.dk/cg, 2002.

[4] 陶跃华，孙茂松. 搜索引擎中相关性反馈技术[J]. 情报理论与实践, 2001(4): 295 - 297.

[5] 张 蕾，侯奠社. 实现概念图工具的若干问题研究[J]. 西安公路交通大学学报, 1998, 18(2): 107 - 110.

[6] Knappe R, Bulskov H, Andreasen T. Similarity Graphs[A]. in Zhong N, Ras Z W, Tsumoto S, Suzuki E. 14th International Symposium on thodologies for Intelligent Systems, ISMIS 2003[C]. Maebashi, Japan: [s.n.], 2003.

[7] 郑实福，刘 挺，秦 兵，等. 自动问答综述[J]. 中文信息学报, 2002, 16(6): 46 - 52.

[8] Na Seung - Hoon, Kang In - Su, Lee Sang - Yool. Question Answering Using a WordNet - based Answer Type Taxonomy [A]. Proceeding of the 11th Text Retrieval Conference[C]. Gaithersburg, USA: [s.n.], 2003.

[9] 刘 群. 计算所汉语词法分析系统[EB/OL]. http://www.ict.ac.cn/freeware, 2003.

(上接第 166 页)

两年一次的参与式设计会议(PDC)在世界各地召开。PDC吸引了研究者、设计者、实践家、工作者和经理人。与会者分享和学习整个设计周期中先进的实践、方法和理论。PD是人们通过一组不同的思想、计划和行动，使他们的工作、技术和社会组织更好地满足人类的需要。PD使用户、项目投资者和其他感兴趣的人能够在形成技术和工作中扮演主要角色，同时这些成果又反应了他们的兴趣。通过PD，世界各地的人们在合作完善技术和社会环境方面，取得了瞩目的成果。相信PD将会有更广阔的应用和发展前景。

参考文献：

[1] Weinberg J B, Stephen M L. Participatory Design in a Human - Computer Interaction Course: Teaching Ethnography Methods to Computer Scientists[A]. UK SIGCSE'02[C]. Leeds: ACM digital library, 2002. 237 - 241.

[2] Preece J, Rogers Y, Sharp H. 交互设计——超越人机交互[M]. 刘晓晖等译. 北京：电子工业出版社, 2003. 198 - 202.

[3] Kukla C D, Binder T, Porter W L, et al, Innovation in Design - Strategies for Designing Together[J]. Tutorials, 1999(15 - 20): 108 - 110.

[4] Rittenbruch M , McEwan G, Ward N, et al. Extreme Participation Moving Extreme Programming Towards Participatory Design[A]. Proceedings of the Participatory Design Conference [C]. Malmo, Sweden: [s.n.], 2002. 23 - 25.

[5] WEBLER T, TULER S, KRUEGER R. What Is a Good Public Participation Process? Five Perspectives from the Public [J]. Environmental Management, 2001, 27(3): 435 - 436.

[6] Cherry C, Macredie R D. The Importance of Context in Information System Design: An Assessment of Participatory Design [J]. Requirements Eng, 1999, 4: 106 - 108.

[7] Garrigou A, Daniellou F, Carballeda G, et al. Activity analysis in participatory design and analysis of participatory design activity [J]. International Journal of Industrial Ergonomics, 1995, 15(1): 313 - 315.

[8] Dix A, Abowd G, Finlay J, et al. Human - computer interaction (second edition)[M]. 北京：电子工业出版社, 2003. 229 - 230.