

模糊算法在信息查询网页中的应用

陈磊松^{1, 2}, 林金山³

- (1. 漳州师范学院 物理与电子信息工程系, 福建 漳州 363000;
2. 厦门大学 信息科学与技术学院, 福建 厦门 361000;
3. 莆田学院 电子信息工程系, 福建 莆田 351100)

摘要: 在模糊逻辑理论的基础上, 将自然语言中的模糊概念用隶属度函数进行描述, 介绍了在浏览器/服务器模式下使用模糊词汇对人员信息进行查询的方法, 作为精确条件查询的一种补充, 更好地反映自然语言的含义, 得到更加精确的结果.

关键词: 模糊算法; 信息查询; 自然语言

中图分类号: TP311 文献标识码: A

Application of fuzzy algorithm in information query over Web pages

CHEN Lei-song^{1, 2}, LIN Jin-shan³

- (1. Dept. of Phy. and Electr. Infor. Eng., Zhangzhou Teachers College, Zhangzhou 363000, China;
2. College of Infor. Sci. and Tech., Xiamen Univ., Xiamen 361000, China;
3. Dept. of Electr. Infor. Eng., Putian College, Putian 351100, China)

Abstract: Based on the theory of fuzzy logic, fuzzy concept in nature language is presented with fuzzy membership functions. A query method of person information with fuzzy nature language words is introduced by browser/server mode. As a supplement for the query of accurate data, it is more suitable in describing the meaning of nature language and achieving more accurate result.

Key words: fuzzy algorithm; information query; nature language

0 引言

自然语言中有许多含义较为模糊的词汇, 如“太小了”、“很高”、“善良”等, 这些模糊的概念很难用传统的数学方法来度量, 尽管如此, 在特定的环境中, 大家用这些模糊概念来描述对象不会引起误解和歧义. 相反, 如果给出的是精确的信息, 如查找有 12368 根头发、4821 根胡子的人, 反而得不到结果. 随着网络数据库技术的发展, 对网络信息检索的要

求越来越高, 传统的精确查询方式已不能满足各种应用领域对数据库查询的要求, 对于精确的查询信息, 容易将其转换为字段条件, 但在使用模糊语言的情况下, 是无法直接转换的. 比如要查询一个比较胖的中等个子的中年男子的信息, “较胖”、“中年”都是模糊概念, 无法用传统的数学方法来度量, 根据常识, 这些描述实际上是一个范围, 而这个范围又没有明确的界限, 与其相邻的状态如“瘦”、“年轻”等, 是交叉的, 因而界限是模糊的. 以男性身高为例, 一般

收稿日期: 2006-06-29

基金项目: 漳州师范学院科研课题(SK05002)

作者简介: 陈磊松(1973-), 男, 福建省福安市人, 漳州师范学院讲师, 厦门大学硕士研究生, 研究方向: 网络与数据库.

将 165 cm ~ 175 cm 左右称为中等身高, 使用传统的程序设计方法, 把小于 165 cm 的身高, 哪怕是 164.9 cm 的身高称为“矮”. 他们只相差 1 mm, 就有中等和矮之分, 这种划分是不恰当的. 目前, 浏览器/服务器 (B/S, Browser/Server) 模式的程序设计方式被越来越广泛地采用. 本文在 B/S 模式下, 将 SQL Server 数据库和 ASP 动态网页技术相结合, 运用模糊算法来设计有关人员信息查询的动态页面, 使一些模糊的自然语言也可以作为查询条件.

1 模糊自然语言的描述

在经典集合中, 某一元素要么属于这个集合, 要么就不属于这个集合, 用这种非此即彼的明晰性来表示集合, 哪怕是很小的变化也可导致落入截然不同的分类中^[1]. 对于一些模糊语言来说, 显然不合理. 经典集合使用的是取值 0, 1 的特征函数来表示事物的归属, 对于模糊性事物, 使用隶属度函数来表示其归属, 隶属度函数是 0 到 1 之间连续变化的逻辑函数. 身高 164.9 cm 与 165 cm 属于同一个集合的程度是很接近的, 如: 168 cm 被认为是“矮”的程度为 0.3, 同时属于“中等”的程度是 0.8.

1.1 模糊集合与隶属函数

解决模糊语言的查询问题, 要引入模糊算法, 将查询条件通过模糊计算, 使用模糊判决方法, 转化为一个查询范围, 进行精确查询^[2]. 首先, 要把模糊语言用模糊集合的方式表示出来.

对于成人信息数据库, 对与模糊语言有关的年龄、身高、体重 3 个字段, 取年龄论域 U_A 为 [10 year, 100 year], 身高论域 U_H 为 [100 cm, 200 cm], 体重论域 U_W 为 [30 kg, 100 kg], 并定义模糊集合如下:

年龄: $S_A = \{ \text{年老, 中年, 年轻} \}$

身高: $S_H = \{ \text{高, 中等, 矮} \}$

体重: $S_W = \{ \text{胖, 中等, 瘦} \}$

对于年龄集合, 定义其隶属度函数如下:

$$\begin{aligned} \text{年老}(u) &= \begin{cases} 0 & 10 \leq u \leq 50 \\ \left[1 + \left(\frac{u-50}{5} \right)^{-2} \right]^{-1} & 50 < u \leq 100 \end{cases} \\ \text{中年}(u) &= \begin{cases} \left[1 + \left(\frac{u-40}{5} \right)^2 \right]^{-1} & 10 \leq u \leq 100 \\ 1 & 10 \leq u \leq 25 \end{cases} \\ \text{年轻}(u) &= \begin{cases} 1 & 10 \leq u \leq 25 \\ \left[1 + \left(\frac{u-25}{5} \right)^2 \right]^{-1} & 25 < u \leq 100 \end{cases} \end{aligned}$$

根据上述定义可以算出, 如果年龄是 30 岁, 则其属于年轻的程度是 0.8, 属于中年的程度是 0.2, 年龄 60 岁属于年老的程度是 0.8. 用类似定义年龄隶属度的方法, 还可以定义成年男性身高、成年女性身高、成年男性体重、成年女性体重的隶属度函数.

如, 成年男性体重的隶属度函数为:

$$\text{胖}(u) = \begin{cases} 0 & 30 \leq u \leq 75 \\ \left[1 + \left(\frac{u-75}{5} \right)^{-2} \right]^{-1} & 75 < u \leq 100 \end{cases}$$

其他的隶属度函数也是类似的, 但需要说明的是, 隶属度函数的确定并不是惟一的, 比如, 对于“胖”的隶属度, 还可以考虑身高的因素, 这些隶属度函数带有一定的主观性, 是凭经验值确定的, 需要在实践中根据反馈的信息进行修正.

1.2 语气算子

语气算子用于表达模糊值的肯定程度. 一种是有强化作用的语气算子, 如“很”、“极”、“相当”等, 使模糊值的隶属度的分布向中央集中, 称集中化算子; 另一种有弱化作用的语气算子, 如“比较”、“稍”等, 使模糊值的隶属度的分布由中央向两边弥散, 称为松散化算子. 用 H_λ 作为语气算子来定量描述模糊值. 若模糊值为 A , 定义: $H_\lambda A = A^\lambda$. 设 H_4 代表“极”或“非常非常”, H_2 代表“很”或“非常”, $H_{1.25}$ 代表“相当”, $H_{0.75}$ 代表“较”, $H_{0.5}$ 代表“有点”或“稍微”, $H_{0.25}$ 代表“稍微有点”.

1.3 模糊化算子

模糊化算子的作用是把词的肯定的意义转化为模糊, 如“大约”、“近乎”等, 如果对数字进行作用, 就把精确数转化为模糊数. 例如身高 160 cm 是精确数, 而“大约 165 cm”就是模糊数; 如果对模糊值进行作用, 就更模糊, 如“年轻”和“近乎年轻”.

用 $E(u, y)$ 表示模糊化算子, 定义:

$$E(u, y) = \begin{cases} e^{-\left(\frac{u-y}{\delta}\right)^2} & |u-y| < \delta \\ 0 & |u-y| \geq \delta \end{cases}$$

如果限制隶属度 ≥ 0.5 且 $\delta = 10$, 对于模糊语言“身高大约 165 cm”, 即 $y = 165$ cm, 根据上述定义容易得到 $156.7 \text{ cm} \leq u \leq 173.3 \text{ cm}$.

1.4 判定化算子

判定化算子与模糊化算子有相反作用, 把模糊值进行肯定化处理, 做出倾向性判断. 如“偏向于”、“多半是”、“倾向于”. 可以用类似“四舍五入”的处理方法, 用隶属度为 0.5 作为分界判断.

例如, 根据以上体重“胖”的隶属度函数定义, 当体重 $u = 80$ kg 时, 其隶属度为 0.5, 则

$$[\text{偏胖}](u) = \begin{cases} 0 & u < 80 \\ 1 & u \geq 80 \end{cases}$$

即当体重超过 80 kg 时, 认为是属于偏胖.

2 模糊自然语言查询的实现

在传统数据库中实现模糊查询, 要对 SQL 语言进行模糊化的扩展^[3,4]. 首先将查询条件通过模糊

计算,转化为一个模糊值的精确范围,再进行精确的SQL查询.按照查询条件,将相应的字段值代入到年龄、身高、体重的隶属度函数中计算出相应的隶属度表,然后设定出总隶属度阈值,再对隶属度查询.

2.1 确定各个模糊值的隶属度

假设查询条件为:相当年轻,身高接近 165 cm 且有点胖的男性.结合上述隶属度函数与语气算子的定义,在使用 SQL Serve 和 ASP 相结合的动态网页设计中,可使用下面的程序生成隶属度表:

```

<%
set conn=Server.CreateObject("ADODB.Connection")
conn.open "dsn=rkxx;uid=sa;pwd="
str1="power(1+power((nl-25)/5.0,2),-1.25)"
str2="exp(-power((sg-165)/10.0,2))"
str3="power(1+power((tz-75)/5.0-2),-0.5)"

```

```

str=str1+" as ua"+ str2+" as uh"+ str3+" as uw"
sql1="insert u_ryxx(lsh, xm, xb, nl, sg, tz) select lsh, xm,
xb, nl, tz from rkxx where xb='男'"
sql21="update u_ryxx set ua=1 where nl<=25"
sql22="update u_ryxx set ua=" + str1 + " where nl>
25"
sql3="update u_ryxx set uh=" + str2
sql41="update u_ryxx set uw=0 where tz<=75"
sql42="update u_ryxx set uw=" + str3 + " where tz>
75"
conn.execute(sql1)
conn.execute(sql21)
.....
%>

```

由以上程序可获得如表 1 所示的隶属度表.

表 1 隶属度表

xm(姓名)	xb(性别)	nl(年龄)	sg(身高)	tz(体重)	U_A (相当年轻)	U_H (接近 165 cm)	U_w (有点胖)
刘道义	男	25	153	62	1	0.236928	0
翁永嘉	男	30	178	75	0.420448	0.184520	0
廖红斌	男	20	160	78	1	0.778801	0.514496
叶辉	男	43	162.5	52.5	0.037059	0.939413	0
陈燊	男	75	170.5	58	0.003123	0.738968	0
沈河林	男	28	172	88	0.680889	0.612626	0.933346

2.2 根据总隶属度查询

根据隶属度表,求出总隶属度,再设置总隶属度的阈值,对隶属度查询.总隶属度为各隶属度中最小的一个,取总隶属度阈值为 0.5,则可用以下程序获得最后结果.

```

<%
set conn=Server.CreateObject("ADODB.Connection")
set res=Server.CreateObject("ADODB.Recordset")
conn.open "dsn=rkxx;uid=sa;pwd="
sql="select * from ryxx where lsh in (select lsh from u_ryxx
where ua>=0.5 and uh>=0.5 and uw>=0.5)"
set res=conn.execute(sql)
%>

```

2.3 实验结果

以上程序是在 Microsoft SQL Server 2000 和 ASP 环境下实现的,能较好地表达使用自然语言进行查询的要求,是对传统精确查询的一种模糊化的扩展,本文中所列出的隶属度函数带有一定的主观性,这也是模糊逻辑的特点^[5],这种主观性是否符合实际,需要在应用中进行一定的修正.

3 结论

在上述的模糊算法中,虽然最后也将模糊概念转换为精确的查询范围,但这与一开始就使用精确

范围是有区别的.如果直接用精确条件替代模糊语言,实际上是直接对模糊语言的近似和简化,是非常粗糙的.而模糊逻辑则是把更多的实际情况考虑在内,用模糊性语言来描述,不是进行简化,而是进行模糊化,使其在一定范围内以渐变的方式实现,可得到更精确的结果.通过模糊扩展 SQL 的方法,容纳了模糊单词、语气算子、模糊化算子和判定化算子,能较好地表达自然语言查询的要求,转化为 SQL 后也易于在动态网页的查询页面设计中实现.

参考文献:

[1] 窦振中.模糊逻辑控制技术及其应用[M].北京:北京航空航天大学出版社,1996.35-68.

[2] 黄波.基于扩展 SQL 的 GIS 模糊查询的表达与实现[J].武汉测绘科技大学学报,1996,21(1):86-89.

[3] Chen Shyi-Ming, Jong Woei-Tzy. Fuzzy query translation for relational database systems[J]. IEEE Transactions on Systems, Man, and Cybernetics-part B: Cybernetics, 1997, 27(4):714-721.

[4] 金林樵.网络数据库技术及应用[M].北京:机械工业出版社,2003.168-228.

[5] 张颖超,叶小岭,吴士芬,等.基于 SQL 的模糊查询技术研究与实现[J].微电子学与计算机,2005,22(1):113-117.