基于模糊概念图的文档聚类及其在 Web 中的应用*

陈 宁[1], 陈 安[2,3], 周龙骧[4], 贾维嘉[5], 罗三定[5]

[1](中国科学院 研究生院 信息学院,北京 100039);
[2](中国科学院 科技政策与管理科学研究所,北京 100080);
[3](中国科学院 软件研究所 软件工程技术研究开发中心,北京 100080);
[4](中国科学院 数学与系统科学研究院,北京 100080);
[5](香港城市大学 计算机科学系,香港 九龙)
E-mail: anchen1@yahoo.com; anchen@otcaix.iscas.ac.cn
http://www.casipm.ac.cn

摘要: 随着 World Wide Web 上数据量的日益庞大,现有的搜索引擎已经不能满足用户日益增长的需求.利用数据挖掘技术,提高搜索效率,实现了查询的用户化.首先提出了模糊概念图的模型来描述词语间的关系,然后在聚类过程中引入概念知识,提出了基于模糊概念图的文档聚类算法,通过分析用户的浏览行为发现兴趣模式.在上述技术的基础上,给出了一种用户化的智能搜索系统的实现策略,通过分析概念间的关系和用户的兴趣模式,评价超链/文档和查询的相关程度,从而帮助用户得到更准确的信息.

关 键 词: 模糊概念图;文档聚类;兴趣模式;用户化智能搜索
中图法分类号: TP393 文献标识码: A

World Wide Web 是由通过超链连接的文档构成的异构的、分布的、动态的数据库.Web 的容量增长迅速,平均每两个小时增加一台服务器,每天增加 100 万个页面.如何从海量信息中发现有用的知识成为知识发现领域所面临的新课题.目前大多数搜索引擎存在着返回结果太多、查询质量低、查询覆盖面小等缺点.随着 Web 信息的急剧增长,现有的搜索引擎已经不能满足用户的信息服务要求.数据挖掘技术可根据用户的个人信息以及以往的浏览行为发现用户的兴趣和偏好,从而为用户提供更准确的查询结果.

本文给出了一种基于模糊概念图的文档聚类和兴趣模式挖掘算法,并介绍了它在 Web 搜索引擎中的应用.鉴于以往的概念树结构不能描述词语之间的复杂关系,本文给出了一种模糊概念图的模型,其中一个概念能够某种隶属程度属于另一个概念.通过把概念知识引入聚类过程,本文提出了基于模糊概念图的文档聚类算法,分析用户的访问记录,发现兴趣模式.在以上技术的基础上,本文提出了一种用户化的智能搜索系统 ICSS,利用模糊概念图和用户的兴趣模式指导搜索,对超链和文档的相关程度进行有效的评价,从而限制了搜索区域,提高了查询效率,实现查询的用户化.本文第 1 节给出了模糊概念图的模型及搜索算法.第 2 节提出了基于模糊概念图的文档聚类和兴趣模式挖掘算法.第 3 节介绍了 ICSS 的搜索策略、超链的优先级和文档的相关度评价规则.最后是结论和进一步的研究方向.

* 收稿日期: 2000-12-06; 修改日期: 2001-06-05
基金项目: 国家自然科学基金资助项目(69983011);中国博士后基金资助项目
作者简介: 陈宁(1974－),女,福建永泰人,博士,主要研究领域为数据挖掘,决策支持系统;陈安(1970－),男,山东东平人,博士,助理研究员,主要研究领域为供应链管理,电子商务;周龙骧(1938－),男,浙江人,研究员,博士生导师,主要研究领域为数据库理论,电子商务;贾维嘉(1955－),男,湖南人,副教授,主要研究领域为网络协议,并行计算;罗三定(1955－),男,湖南人,副教授,主要研究领域为网络安全.

1 模糊概念图

概念层次是一种在数据挖掘中广泛应用的领域知识.面向属性的概念树提升算法[1]发现关系数据库的特征规则、区别规则和数量规则.Ramakrishnan 等人[2]利用数据项的层次关系发现不同概念层次上的数据项之间的扩展关联规则.Han 等人[3]提出了一种对概念树从上到下逐层编码的算法,发现多层次的关联规则.以往的这些概念层次树针对交易数据库的数据项集合或者某个属性的值域,概念之间的层次清晰,一个概念或者属于或者不属于另一个概念,一个结点只能有一个父结点.而词语之间的关系复杂,不能用概念树这种简单的结构来描述.首先,词语在概念上的关系不是绝对的属于或不属于,一个词可以以某种隶属程度属于另一个词.其次,对词语而言强制只有一个父概念是不符合实际情况的.因此,概念树描述能力的局限性使它不能表达复杂的关系.我们在定义上对其进行了扩展,给出了一种模糊概念图的模型.如图 1 和图 2 所示.

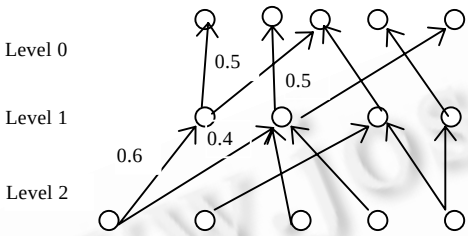


Fig.1 Fuzzy concept graph
图 1 模糊概念图

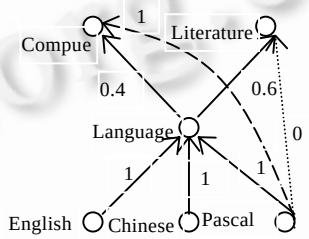


Fig.2: Arc, hyper-arc and anti-arc
图 2 弧、超弧和反超弧

定义 **1**. 令 C 是一组概念,我们按照概念的外延把 C 划分为不同的层次,构成一个有向的、带权值的、非循环的、无重边的多分图,称为模糊概念图.其中每个结点表示一个概念,最上层的结点代表最一般的概念,最下层的结点代表最具体的概念,结点 p 所在的概念层记作 level(p).弧反映两个相邻层次上结点之间的隶属关系,权值表示隶属度.一般地,模糊概念图 $G=\{V,E\}$可以表示为

(1) $V(G)=\{p|p\in C\}$;

(2) $E(G)=\{\langle p,q\rangle|0<\text{weight}(p,q)\le 1,\text{level}(q)=\text{level}(p)-1\}$;

(3) $\forall p\in V(G),\ \sum_{\substack{\text{level}(q)=\text{level}(p)-1\\ \langle p,q\rangle\in E(G)}}\text{weight}(p,q)\le 1$.

定义 **2**. 令 $p,q\in V(G)$,level(p)>level(q),如果存在一条链 $c=\langle x_1,x_2,\ldots,x_k\rangle,x_1=p,x_k=q,x_i\in V(G),\langle x_i,x_{i+1}\rangle\in E(G)$,则称 p 在概念上隶属于 q,q 是 p 的祖先,p 是 q 的后代.p 通过 c 相对于 q 的隶属度 $\text{belong}(p,q,c)=\prod_{i=1}^{k-1}\text{weight}(x_i,x_{i+1})$.特别地,如果$\langle p,q\rangle\in E(G)$,则称 q 是 p 的父概念,p 是 q 的子概念,belong(p,q,c)=weight(p,q).

定义 **3**. 令 $p,q\in V(G)$,level(p)>level(q),则 p 相对于 q 的隶属度定义为

$$\text{belong}(p,q)=\begin{cases}\text{weight}(p,q), & \text{if } \langle p,q\rangle\in E(G),\\ \sum_{\substack{c=x_1,x_2,\ldots,x_k,\\ x_1=p,x_k=q,\\ \langle x_i,x_{i+1}\rangle\in E(G)}}\text{belong}(p,q,c), & \text{Otherwise.}\end{cases}$$

定理 **1**. 在模糊概念图中,结点 p 相对于任意一层上的所有结点的隶属度之和不大于 1,即

$$\forall p\in V(G),\forall i\ge 1,\ \sum_{\text{level}(q)=\text{level}(p)-i}\text{belong}(p,q)\le 1.$$

证明:令 level(p)=k. $\sum_{\text{level}(q)=k-1}\text{belong}(p,q)\le\sum_{\text{level}(q)=k-1}\text{weight}(p,q)\le 1$.

假设 $m>1$, $\sum_{\text{level}(q)=k-m+1}\text{belong}(p,q)\le 1$ 成立.那么,

$$\sum_{\text{level}(q)=k-m}\text{belong}(p,q)=\sum_{\text{level}(q)=k-m}\sum_{\text{level}(r)=k-m+1}\text{belong}(p,r)\times\text{belong}(r,q)$$

$$= \sum_{\text{level}(r)=k-m+1} \text{belong}(p,r) \times \Big(\sum_{\text{level}(q)=k-m} \text{belong}(r,q) \Big)$$
$$\le \sum_{\text{level}(r)=k-m+1} \text{belong}(p,r) \le 1.$$

所以, $\forall i \ge 1, \sum_{\text{level}(q)=\text{level}(p)-i} \text{belong}(p,q) \le 1$. □

推论 **1**. $\forall p,q \in V(G)$,level(p)>level(q),有 belong(p,q)≤1.

证明:显然,belong(p,q)≤ $\sum_{\text{level}(r)=\text{level}(q)} \text{belong}(p,r) \le 1$. □

概念在含义上的复杂性可能造成概念的层次不清晰,一个概念可能直接隶属于不同层次上的多个概念.为此,我们用超弧(hyper-arc)来表示跨越两个或以上层次的两个概念之间的关系,记作 $E'(G)=\{\langle p,q\rangle | \text{weight}(p,q) \in [0,1], \text{level}(q)<\text{level}(p)-1\}$.

定义 **4**. 令 $p,q \in V(G)$,level(p)>level(q),则 p 相对于 q 的隶属度定义为

$$\text{belong}(p,q)=\begin{cases} \text{weight}(p,q), & \text{if } \langle p,q\rangle \in E(G) \cup E'(G), \\ \sum_{\substack{c=x_1,x_2,\ldots,x_k,\\ x_1=p,x_k=q,\\ \langle x_i,x_{i+1}\rangle \in E(G)}} \text{belong}(p,q,c), & \text{Otherwise.} \end{cases}$$

定理 **2**. 在扩展模糊概念图中,$\forall p,q \in V(G)$,level(p)>level(q),有 belong(p,q)≤1.

证明:令 level(p)=k.当 level(q)=k−1 时,belong(p,q)≤weight(p,q)≤1.

假设 $r \in V(G)$,level(r)=k−m+1,有 belong(p,r)≤1.那么,$q \in V(G)$,level(q)=k−m,如果 $\langle p,q\rangle \in E'(G)$,则 belong($p$,$q$)=weight($p$,$q$)≤1.否则,

$$\text{belong}(p,q) = \sum_{\text{level}(r)=k-m+1} \text{belong}(p,r) \times \text{belong}(r,q) \le \sum_{\text{level}(r)=k-m+1} \text{belong}(r,q) \le 1.$$

所以,$\forall p,q \in V(G)$,level(p)>level(q),belong(p,q)≤1. □

为了实现对概念图的搜索,需要存储每个结点的概念层次、父概念和子概念、弧的类型、权值等信息.Procedure Ancestor(p,G,max_depth,min_belong)利用广度优先搜索,由下到上逐层扩展得到 p 的祖先和隶属度,存储在数组 ancestor 中,参数 max_depth 用来限制搜索的深度,min_belong 是最小隶属度.类似地,Procedure Descendant(p,G,max_depth,min_belong)得到 p 的后代和隶属度.

```
Struct vertice (p){
  level;//the level number of p
  num_parent;//outdegree of p
  num_child;//indegree of p
  parent [num_parent]//the information of arcs pointed from p
    {q;   //one parent of p
     type;//0:arc,1:hyper-arc
     weight (p,q)//the belongingness of p with respect to q}
  child [num_child]//the information of arcs pointed to p
    {r;   //one child of p
     type;//0:arc,1:hyper-arc
     weight (r,p)//the belongingness of r with respect to p}}
```

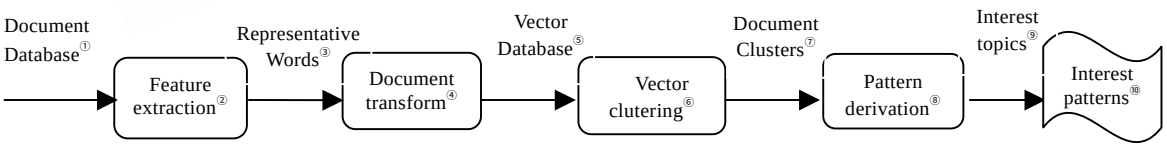
上述数据结构用来存储结点 p 所在的概念层次、p 的父概念和子概念、弧的类型、权值等信息.Procedure Ancestor (p,G,max_depth,min_belong)利用广度优先搜索,由下到上逐层扩展得到 p 的祖先和隶属度,参数 max_depth 用来限制搜索的深度,min_belong 是最小隶属度.搜索得到的祖先结点及隶属度等信息存储在数组 ancestor 中.类似地,Procedure Descendant (p,G,max_depth,min_belong)得到 p 的后代和隶属度.

```
Procedure Ancestor (p,G,max_depth,min_belong)
        //Find the ancestors and of p from G via a chain no longer than max_depth and the belongingness
          no less than min_belong
Input: Concept graph G={v1,v2,…,vvernum}, where Vernum is the number of vertex in G;
```

```
        p: a vertice in G;
        max_depth: a parameter to control the depth of search;
        min_belong: a parameter to control the minimum belongingness of p with respective to its ancestor;
Output: ancestor_num;//the number of ancestors of p;
        Ancestor [ancestor_num];//array of ancestors;
Begin
      Belong (p,p)=1.0; Seeds=Empty; ancestor_num=0;
      For i:=1 to Vernum do {visited [v_i]=0; belong (p,v_i)=0;}
      Seeds.append (p);
      While Seeds<> Empty do
       {Curr:=Seeds.first();//pop up the first element of Seeds
          If   level (Curr)>level (p)−max_level and Curr<>p then
          {ancestor_num++;
           ancestor [ancestor_num]=Curr;
           For i=1 to Curr. num_parent;//find all parents of Curr
            {v=Curr.parent[i].q;
                    If not Curr.parent[i].type then
                {belong (p,v)=belong (p,v)+Curr.parent[i].weight (Curr,v)×belong (p,Curr);
                 visited[v]=1;
                }
               If not visited [v] then Seeds.append(v);
            }//End For
          }//End If
      Seeds.delete (Curr);
      }//End While
      For i=1 to p.num_parent;    //find the hyper-arc and anti-arc pointed from p
      If p.parent[i].type<>0 then
          {v=p.parent[i].q;
          belong(p,v)=weight(p,v);//reassign the belongingness if hyper-arc or anti-arc exists
       }
      For (i=1;i≤ancestor_num;i++)
      If belong (p,ancestor[i])<min_belong then
      {delete(ancestor[i]);ancestor_num--;}
End
```

2 CDCG——基于模糊概念图的文档聚类及兴趣模式的挖掘算法

Web 服务器自动监控用户的日常活动,将每次访问的时间、用户的网络地址、目的信息的网络地址,传输的信息量存储在访问日志中,为文档聚类提供了数据源.以往的文档聚类算法[4,5]把每个文档转换为一个带权矢量,其中每个词作为一个属性,词的出现频率为属性值,然后对矢量进行聚类.这种方法是建立在属性正交,即各属性互相独立的假设之上的.事实上,这种假设是不成立的.本文提出了一种基于模糊概念图的聚类算法,首先对文档进行预处理,过滤掉大量的虚词,合并意义相近的词,计算其余的词在每个文档中的重要程度,抽取重要程度高的词组成代表词汇,然后根据模糊概念图中词之间的关系将文档转换为矢量,对矢量进行聚类,最终得到一组兴趣模式.如图 3 所示.



①文档数据库,②特征抽取,③代表词汇,④文档转换,⑤矢量数据库,⑥矢量聚类,⑦文档聚类,⑧模式提取,⑨兴趣主题,⑩兴趣

Fig.1 Mining interest patterns

图 1 兴趣模式的挖掘过程

2.1 特征抽取

首先删除出现频率很高但很少与文档的内容有关的虚词,同时合并具有相同词根且含义相近的词;然后利用 FTIDF(frequency term-inverse document frequency)方法[6,7]计算词在每个文档 d_i 中的重要程度;根据阈值 T 选择一组词 W_i,满足$\forall w \in W_i, tfidf(w,d_i) \geq T$;合并抽取的词汇 $W=\bigcup_{i=1}^{n} W_i$ 作为代表词汇.

2.2 文档的矢量转换

一个词对文档的重要性不仅取决于它本身的出现频率,还取决于它的子概念的出现频率以及子概念的隶属度.给定文档 d 和词 w,w 在 d 中出现的次数称为绝对频率,记作 $val(w,d)$.w 对祖先 v 的相对频率等于β*$val(d,w)$*belong(w,v),其中$\beta \in [0,1]$是衰减因子.词的频率等于绝对频率和相对频率之和.矢量转换以 W 及其祖先为属性集合,属性在文档中出现的频率为属性值,经过标准化后映射为一个矢量,从而把文档数据库 D 转换为矢量数据库 D'.

Procedure Transform (W,G,D,D')
Input: D:document database, W:the set of representative words of D, G:fuzzy concept graph;
Output: D' vector database transformed from D;
Begin
  $A=\varnothing$; //the set of attributes of the transformed database;
  For all $w \in W$ do
  {Perform procedure *Ancestor*(w,G,max_depth,min_belong); //get the ancestors and belongingness of w;
   $A=A \cup ancestor(w)$;
  }
  For each document $d \in D$ do
  {For each $A_i \in A$ do Calculate $val(A_i,d)$;
    For each $A_i \in A$ do $fre(A_i,d)=val(A_i,d)$;
    For each $w \in W$ do
      {For each $v \in ancestor(w)$ do
        $fre(v,d)=fre(v,d)+\beta * val(w,d) * belong(w,v)$; //add the relative frequency to the absolute frequency;
      }// Endfor
   Normalize the $vector(fre(A_i,d)|A_i \in A)$ to d';
  }//Endfor
  Set A as the set of attributes of D';
  Map d to a normalized vector d' of D';
End

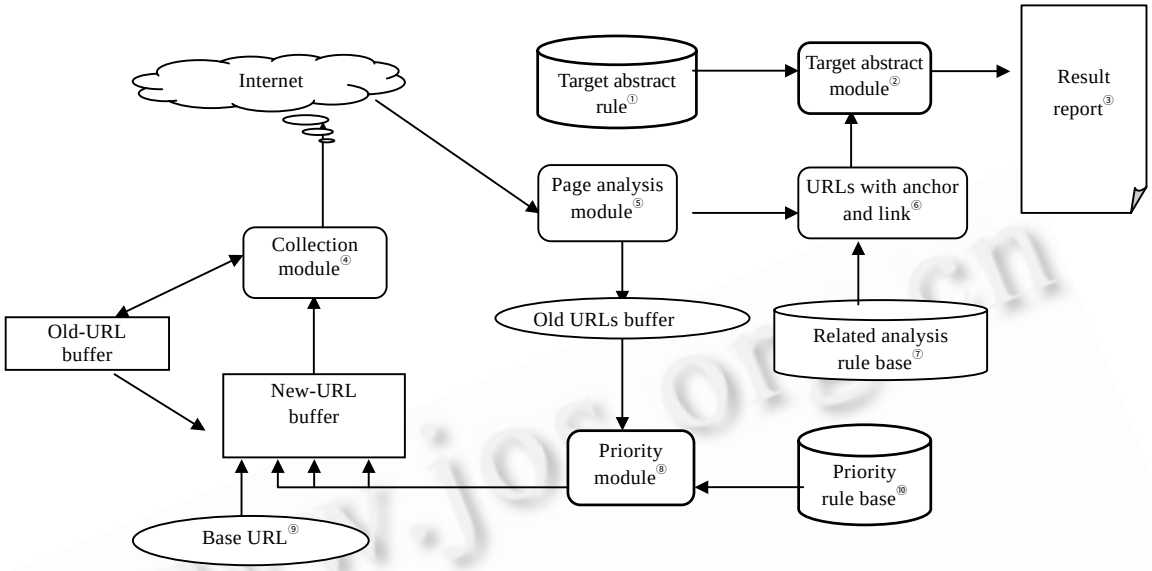
2.3 文档矢量聚类及模式抽取

令 $D'=\{X_1,X_2,\ldots,X_n\},A_1,A_2,\ldots,A_m$ 为 D'的一组属性.从直观上看,两个文档包含的公共词越多,出现的频率越高,它们的相似程度就越高,距离也就越小,因此两个文档之间的距离用矢量的 Jaccard 系数$(X,Y \in D', d(X,Y)=1-\sum_{i=1}^{m} x_i y_i \Big/ \left(\sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} y_i^2 - \sum_{i=1}^{m} x_i y_i\right)$来衡量.采用模糊 K-Means 算法[8,9],对文档进行聚类.首先随机地选取 k 个初始聚类中心,计算相应的隶属矩阵;然后修改当前的聚类中心,并重新计算隶属矩阵;如果新的聚类的目标函数值小于原先的聚类,则用新聚类代替原聚类;重复执行这一过程直至目标函数值不再减小为止.迭代过程结束时得到的中心矩阵和隶属矩阵分别记作 Z^*,W^*.如果对象 X_i 与 Z_l^* 的距离不小于 Z_l^* 与其他聚类中心的最小距离或者 X_i 对 Z_l^* 的隶属度不大于最小阈值λ,则从 Z_l^* 对应的聚类中删除 X_i,并重新计算聚类中心.每个聚类对应一个兴趣模式,用聚类的中心矢量表示.每个中心由一组反映用户的兴趣主题的词构成,作为用户的兴趣模式.

3 ICSS——用户化的智能搜索系统

目前的搜索引擎大多基于关键字的用户界面,限制了查询语言的表达能力,返回的结果太多,但许多不是用户所需要的,而且,查询缺乏用户化.针对目前多数搜索引擎存在的搜索结果质量不高、低质无效链接很多的缺

陷,我们致力于研究设计一种能够提高信息搜索质量的智能搜索系统[10].如图 4 所示.ICSS 搜索引擎接受查询语句后,反复执行下述步骤,直到 New-URL buffer 为空.



①目标抽取规则库,②目标抽取模块,③查询结果,④信息收集模块,⑤页面分析模块,
⑥相关分析模块,⑦相关分析规则库,⑧优先级模块,⑨基地址,⑩优先级规则库.

Fig.2 Search mechanism

图 2 搜索机制

(1) 搜索引擎首先得到一组初始优先级的基地址,存入 New-URL buffer 中.New-URL buffer 是多优先级堆栈,采用优先级优先及后进先出的策略.Old-URL buffer 保存所有在本次搜索过程中访问过的 URL,以避免重复访问.

(2) 收集模块负责从因特网上抽取 Web 文档.对每个新取出的 URL,首先查找 Old-URL buffer,如果该 URL 已经被访问过,就放弃请求.为了防止超时等待,如果访问页面的时间超过某一阈值,系统自动放弃请求.

(3) 超链是 HTML 文档的主要组成元素,是访问其他页的途径.超链由两部分构成:链源(anchor)通常是对目的页面内容的概括,地址(link)是页面的路径和名称.页面分析模块从已获取的页面中抽取所有超链的地址和链源.

(4) 优先级模块根据超链与用户的查询及兴趣的匹配程度计算优先级,把它们放到 New-URL buffer 中.

(5) 相关分析模块分析文档和用户查询和兴趣的匹配,评价其相关程度.

(6) 目标抽取模块从查询得到的相关页中抽取目标信息,输出到结果列表中.

3.1 搜索路径的优先级评价

当用户面对一组超链时,通常根据经验判断超链的相关性,首先选择与查询目的最相关的超链,如果没有找到所需要的信息,再依次浏览次相关的页.ICSS 的搜索引擎模拟人工搜索的策略,根据热点词汇和超链(HL)的链源及地址的匹配程度评价超链与用户查询要求的相关程度.相关程度越高,搜索的优先级也越高.对于只含有一个关键字的简单查询语句 q,热点词汇不仅包括查询关键字 W_q 和用户兴趣模式的关键字 W_u ,还包括这些关键字的同义概念及后代概念.W_q 经过概念扩展后得到 $\overline{W_q}$,W_u 概念扩展为 $\overline{W_u}$.如果一个链源包含 $\overline{W_q}$ 的词,那么该链源指向的内容很可能与查询相关;如果一个链源包含 $\overline{W_u}$ 的词,那么该链源指向的内容很可能是用户感兴趣的.对于搜索过程中得到的一个超链 HL,令 A 为包含在链源中的有意义的关键字的集合,L 为包含在地址中的有意义的关键字的集合.

$$G(A,q)=\begin{cases}1, & \overline{W_q}\cap A\neq\varnothing\\ 0, & \text{Otherwise}\end{cases};\quad G(A,u)=\begin{cases}1, & \overline{W_u}\cap A\neq\varnothing\\ 0, & \text{Otherwise}\end{cases};$$

$$G(L,q)=\begin{cases}1, & \overline{W_q}\cap L\neq\varnothing\\ 0, & \text{Otherwise}\end{cases};\quad G(L,u)=\begin{cases}1, & \overline{W_u}\cap L\neq\varnothing\\ 0, & \text{Otherwise}\end{cases}.$$

复合查询语句通常由关键字和 and,or,not 等逻辑运算符构成,优先级评价方法如下:

Case 1: $q=q^1$ and q^2: $G(A,q)=G(A,q^1)\wedge G(A,q^2)$;$G(L,q)=G(L,q^1)\wedge G(L,q^2)$;

Case 2: $q=q^1$ or q^2: $G(A,q)=G(A,q^1)\vee G(A,q^2)$;$G(L,q)=G(L,q^1)\vee G(L,q^2)$;

Case 3: q=not q^1: $G(A,q)=\neg G(A,q^1)$;$G(L,q)=\neg G(L,q^1)$;

则
$$G(A,q,u)=G(A,q)\times G(A,u);G(L,q,u)=G(L,q)\times G(L,u),$$
$$G(HL,q,u)=G(A,q,u)\vee G(L,q,u);G(HL,q)=G(A,q)\vee G(L,q).$$

因此,超链 HL 的链源对用户 u 的查询 q 的优先级评价规则为

If $G(HL,q,u)$ then $P=P+1$;

Else If $\neg G(HL,q)$ then $P=P-1$;

当用户提出一个查询请求时,搜索引擎从一组基地址开始搜索,并给基地址赋予一个初始优先级.搜索引擎每次处理一个超链 HL,如果 $G(HL,q,u)$等于 1,说明 HL 既与查询目的有关,又与用户兴趣有关,则 HL 的优先级加 1;如果 $G(HL,q)$等于 0,说明 HL 与查询目的无关,则 HL 的优先级减 1;否则 HL 的优先级不变.在查询的过程中,如果某个超链的优先级降低为 0,就停止扩展.显然,从某个初始优先级的地址开始的搜索经过若干次优先级匹配失败后,最终会自动终止,从而限制了搜索的区域,避免了无效的搜索.

3.2 文档的相关分析

文档的相关分析是一个分析文档和用户查询的匹配.评价其相关程度的过程.搜索引擎的相关性判断能力越强,对用户的帮助就越大.以往的搜索引擎通常根据关键字在文档中出现的频率判定文档与查询的相关程度,关键字出现的越多、越早,相关程度也就越大,也就是说,文档对查询的相关程度仅仅由查询关键字在文档中出现的频率决定,而忽略了不出现在查询语句中但与查询关键字概念相关的词以及提出查询请求的用户.利用模糊概念图和用户的兴趣模式,搜索引擎可以实现查询的用户化.假设 q 是用户 u 提出的一个单关键字查询语句,w 是出现在文档 d 中的词,那么 w 对 q 和 u 的相关度由 w 相对于 W_q 和 W_u 中关键字的隶属程度决定:

$$\text{weight}(w,q)=\begin{cases}\max\limits_{p\in W_q}\text{belong}(w,p), & w\in\overline{W_q},\\ 0, & \text{Otherwise},\end{cases}$$

$$\text{weight}(w,u)=\begin{cases}\max\limits_{p\in W_u}\text{belong}(w,p), & w\in\overline{W_u},\\ 0, & \text{Otherwise},\end{cases}$$

$$\text{weight}(w,q,u)=\text{weight}(w,q)\times\text{weight}(w,u).$$

文档 d 对查询 q 和用户 u 的相关度 $Relevance(d,q,u)=\sum_{w\in d}\text{weight}(w,q,u)\times\text{frequency}(w,d)$.

对于复合查询 q 的相关评价方法如下:

Case 1: $q=q^1$ and q^2: $Relevance(d,q,u)=\min(Relevance(d,q^1,u),Relevance(d,q^2,u))$;

Case 2: $q=q^1$ or q^2: $Relevance(d,q,u)=\max(Relevance(d,q^1,u),Relevance(d,q^2,u))$;

Case 3: $q=q^1$ not q^2: $Relevance(d,q,u)=Relevance(d,q^1,u)-Relevance(d,q^2,u)$.

4 结 论

计算机网络的普及使 Internet 成为世界上最大的信息网,为了快速、高效地找到网上的知识,研究 Web 上的挖掘成为近期数据挖掘的重要课题.本文提出了模糊概念图的模型以及基于模糊概念图的文档聚类算法,发现用户的兴趣模式.在此基础上,给出了一种智能搜索系统的搜索策略,用概念关系和兴趣模式评价网页内容和搜索链的相关度,从而提高查询的效率,实现查询的用户化.下一步的工作包括研究单词术语的自动分类、模糊概念图的模型优化、查询语言的语义扩充以及符合信息需求的模糊数学描述等.

References:

[1] Han, J., Cai, Y., Cercone, N. Knowledge discovery in databases: an attribute-oriented approach. In: Yuan, Le-yan, ed. Proceedings of the 18th International Conference on Very Large Data Bases. Vancouver: Morgan Kaufmann, 1992. 547~559.

[2] Srikant, R., Agrawal, R. Mining generalized association rules. In: Umeshwar, D., Gray, P.M.D., Shojiro, N., eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich: Morgan Kaufmann, 1995. 407~419.

[3] Han, J., Fu, Y. Discovery of multiple-level association rules from large database. In: Umeshwar, D., Gray, P.M.D., Shojiro, N., eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich: Morgan Kaufmann, 1995. 420~431.

[4] Oren, Z., Oren, E., Omid, M., *et al*. Fast and intuitive clustering of web document. In: Heckerman, D., Mannila, H., Pregibon, D., eds. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'97). Newport Beach, CA: AAAI Press, 1997. 287~290.

[5] Cheung, D.W., Kao, B., Lee, J. W. Discovering user access patterns on the world-wide-web. In: Lu Hong-jun, Motoda, H., Liu, Huan, eds. Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore: World Scientific, 1997. 303~316.

[6] Salton, G., Buckley, C. Term-Weighting approaches in automatic text retrieval. Information Processing and Management, 1988,24(5):513~523.

[7] Oren, Z. Clustering web documents: a phrase-based method for grouping search engine results [Ph.D. Thesis]. Seattle, WA: University of Washington, 1999.

[8] Bezedek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.

[9] Ruspini, E.H. A new approach to clustering. Information Control, 1969,19(15):22~32.

[10] Luo, San-ding. Efficient intelligent search system for web information mining (EIS). In: Goscinski, A., Horace, H.S.I, Jia, Wei-jia, *et al*, eds. Proceedings of the 4th International Conference on Algorithms and Architecture for Parallel Processing (ICA3PP 2000). Hong Kong: World Scientific Publishing, 2000. 716~717.

A Documental Clustering Algorithm Based on Fuzzy Concept Graph and Its Application in Web*

CHEN Ning[1], CHEN An[2,3], ZHOU Long-xiang[4], JIA Wei-jia[5], LUO San-ding[5]

[1](*School of Information Sciences and Engineering, Graduate School, The Chinese Academy of Sciences, Beijing* 100039, *China*);

[2](*Institute of Policy and Management, The Chinese Academy of Sciences, Beijing* 100080, *China*);

[3](*Technology Center of Software Engineering, Institute of Software, The Chinese Academy of Sciences, Beijing* 100080, *China*);

[4](*Academy of Mathematics and System Sciences, The Chinese Academy of Sciences, Beijing* 100080, *China*);

[5](*Department of Computer Sciences, City University of HongKong, HongKong, China*)

E-mail: anchen1@yahoo.com; anchen@otcaix.iscas.ac.cn

http://www.casipm.ac.cn

Abstract: With the explosive growth of data available on World Wide Web, it seems that the current search engines cannot meet the increasing requirement of users. This paper focuses on improving the effectiveness and the efficiency of Web search with data mining technology. A documental clustering algorithm is presented integrated with fuzzy concept graph for mining interest patterns. Based on the above technology, an intelligent customized search system is proposed that enables users to obtain useful information according to the relation of concepts and own interests. The strategy is to evaluate the relevance of documents effectively based on fuzzy concept graph and user's personal interests.

Key words: fuzzy concept graph; documental clustering; interest pattern; customized search

* Received December 6, 2000; accepted June 5, 2001
Supported by the National Natural Science Foundation of China under Grant No.69983011; the Postdoc. Research Fund of China