基于模糊文本分类的多知识领域专家推荐方法

李 明 刘 鲁 王 君 黄兆东

(北京航空航天大学 经济管理学院，北京 100191) (北京航空航天大学 工程系统工程系，北京 100191)

摘 要：在知识管理系统中，推荐满足知识需求的专家是隐性知识共享的有效手段. 为了推荐多知识领域的专家，以及完整的识别用户的知识需求，提出了基于模糊文本分类的多知识领域专家推荐方法. 首先建立模糊文本分类器，通过对专家上传的文档进行模糊文本分类，结合时间等因素建立专家的知识模型. 通过分析浏览文档日志识别用户直接知识需求，在此基础上采用信息熵计算用户的潜在的知识需求. 然后根据专家的知识模型与用户的知识需求模型的匹配度进行专家推荐. 通过在知识管理系统中的成功应用验证了该方法的可行性与有效性.

关 键 词：文本处理；专家系统；基于知识的系统；知识管理

中图分类号：TP 18

文献标识码：A **文 章 编 号**：1001-5965(2009)10-1254-04

Approach to expert recommendation with multiple knowledge areas based on fuzzy text categorization

Li Ming Liu Lu Wang Jun

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Huang Zhaodong

(Dept. of System Engineering of Engineering Technology, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Recommending an appropriate expert in knowledge management systems is an effective and efficient way to utilize tacit knowledge. To recommend the experts with multiple knowledge areas and identify the user's knowledge needs completely, an approach to expert recommendation with multiple knowledge areas based on fuzzy text categorization was proposed. Firstly, fuzzy text classifier was constructed. An expert profile was built by classifying the newly registered knowledge artifacts fuzzily and with volume and time factors. Knowledge needs model was composed of explicit knowledge needs and implicit knowledge needs. Explicit knowledge needs was identified by analyzing the browsing logs. Implicit knowledge needs was measured by the entropy of the explicit knowledge needs. Expert was recommended based on the matching degree of expert profile and the user's knowledge needs model. The proposed approach was developed and was used successfully in the knowledge management system. The approach was proved to be applicable.

Key words: text processing; expert systems; knowledge based systems; knowledge management

有效的知识共享是企业保持持续竞争优势的重要手段，其关键是对构成企业核心能力的隐性知识的共享. 而隐性知识存储于人们的头脑中，难以进行编码化存储，因此通过向专家咨询是共享隐形知识非常有效的方法. 随着企业规模的增大，不同部门之间的交流越来越少，确定"'谁'知道'什么'"的问题也变得越来越困难，从而对于企业来说有效地进行专家识别与推荐是进行隐形知识

收稿日期：2008-11-30

基金项目：国家自然科学基金资助项目(70671007，70871006)；教育部博士点基金资助项目(200800060005)

作者简介：李 明(1981—)，男，山东枣庄人，博士生，limigsd@gmail.com.

共享的挑战[1].

很多企业都建立了知识管理系统,以提高企业的知识管理水平.传统的知识管理系统中的专家模型是构建在专家数据库之上的,其中的专家数据是由专门的员工来维护的,具有很大的主观性和盲目性,并且不能及时地对专家的信息进行更新[2].目前国内外对各种自动建模方法的研究已经取得了一定的成果.文献[2]提出了使用文本分类的方法进行专家模型维护,该方法通过专家建模,将专家归到某一具体类别中,用户根据分类进行专家检索.然而对于企业中的员工,特别是新员工来说,经常面临难以明确自己的知识需求,即使提供了专家目录,也无法确定需要咨询的专家的问题.文献[3]提出了基于概率语言模型的专家推荐方法,通过对专家文本形式的表示,根据检索条件进行专家的推荐.文献[4]以专家为关注点,将两种概率语言模型结合进行专家建模.文献[5]对两种概率语言模型进行一般化建模,提出了统一的基于概率语言模型的专家推荐方法.文献[2—5]的方法克服了传统数据库方法的弊端,但均假设用户的知识需求领域与专家的知识领域是单一的,而用户的知识需求与专家经常是多知识领域的.此外,通过当前文档检索识别用户知识需求的方法无法识别用户的潜在知识需求.譬如为了比较各类发动机的特点,用户检索并浏览了涡扇发动机、涡喷发动机与涡桨发动机等相关文档,其直接知识需求为涡扇发动机、涡喷发动机与涡桨发动机等知识领域,而其潜在的知识需求是较为综合的动力装置知识领域.

为了推荐多知识领域的专家和完整的识别用户的知识需求,本文提出了基于模糊文本分类的多知识领域专家推荐方法,在建立专家知识模型与用户知识需求模型的基础上,通过二者的匹配度进行专家推荐.

1　多知识领域的专家推荐过程

专家推荐包括 5 个步骤:构造模糊文本分类器,构造专家的知识模型,构造用户的知识需求模型,计算专家知识模型与用户知识需求模型的匹配度,专家推荐,如图 1 所示.

通过构造模糊文本分类器,对专家上传到知识库中的文档进行模糊文本分类,结合数量、时间等因素建立专家知识模型,然后更新专家模型库.通过分析用户一段时间内的文档浏览日志建立用户的知识需求模型,然后更新知识需求库.最后根据专家知识模型与用户知识需求模型的匹配度进行多知识领域专家推荐.

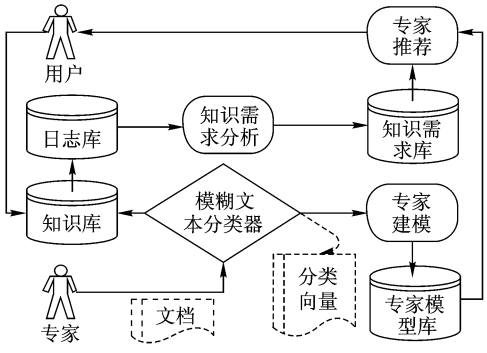


图 1　多知识领域的专家推荐过程

2　模糊文本分类器构建

构建模糊文本分类器的目的是对进入知识库的文档进行模糊文本分类.模糊文本分类器是由各知识领域的特征向量构成.

2.1　准备领域本体和训练文档集

领域本体是对概念及概念之间关系的定义.领域本体通常表现为一棵概念化的分类树,树的上层是与领域独立的比较广泛的概念,越往树的下方越具有领域相关性.使用领域本体可以澄清该领域的知识结构.领域本体中每个节点作为一个知识领域.

图 2 是航空领域局部领域本体.

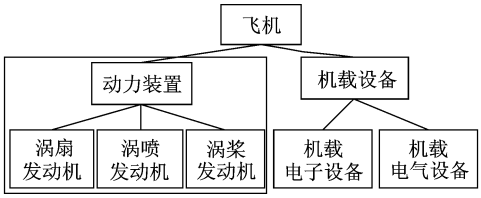


图 2　航空领域局部领域本体

训练文档集是已经准确划分到各知识领域的文档.在进行使用前需要对训练文档进行预处理,包括中文分词,去掉介词、形容词等.

2.2　计算训练文档集中特征词词频

特征词是指在文档中具有解释能力或区分能力的词,特征词构成特征词集合 $T=\{t_1,\cdots,t_q\}$.知识领域 j 中训练文档的集合表示为 $D_j=\{D_{j1},D_{j2},\cdots,D_{jd}\}$.$d$ 是在知识领域 j 中训练文档的数量.

2.3　构造知识领域特征向量

知识领域 j 的特征向量 $\boldsymbol{F}_j^v$ 是由特征词以及该特征词在知识领域 j 的权重构成:

$$\boldsymbol{F}_j^v=\{(t_1,w_{j1}^f),(t_2,w_{j2}^f),\cdots,(t_q,w_{jq}^f)\}$$

其中,q 是特征向量的维度,也是特征词的数量;w_{jq}^f 是特征词 t_q 在知识领域 j 中的权重;f 表示 w_{jq}^f 是知识领域特征向量的权重.

$$w_{jq}^{f}=n_{jq}^{tf}\cdot n_{q}^{idf} \tag{1}$$

其中，n_{jq}^{tf} 表示特征词 t_q 在知识领域 j 出现的频率，n_{jq}^{tf} 值越大，表示特征词 t_q 对知识领域 j 越重要；n_{q}^{idf} 表示特征词 t_q 的反文档频率，含有特征词 t_q 的知识领域数越少，n_{q}^{idf} 值越大，特征词 t_q 对含有该特征词的知识领域的区分度就越高.

$$n_{jq}^{tf}=n_{jq}^{\mathrm{freq}}\Big/\sum_{k=1}^{q}n_{jk}^{\mathrm{freq}} \tag{2}$$

其中，n_{jq}^{freq} 是特征词 t_q 在知识领域 j 出现次数.

$$n_{q}^{idf}=\log(N/n_q) \tag{3}$$

其中，N 是知识领域的数量；n_q 是含有特征词 t_q 的知识领域的数量.

3 基于模糊文本分类的专家建模

专家建模是对专家的知识领域以及在该领域的知识水平进行建模. 通过对专家上传的文档进行模糊文本分类，然后结合数量、时间等因素更新专家模型.

3.1 模糊文本分类

通过模糊文本分类，确定文档与各知识领域的相关度，从而获得文档的分类向量. 文档 j 的分类向量表示为

$$\boldsymbol{D}_j^v=\{(k_1,w_{j1}^d),(k_2,w_{j2}^d),\cdots,(k_N,w_{jN}^d)\}$$

其中，N 是知识领域的数量；k_N 表示知识领域 N；w_{ji}^d 是文档 j 与知识领域 i 的相关度；d 表示 w_{ji}^d 是文本分类特征向量的权重；w_{ji}^d 为知识领域特征向量与文档向量的内积，表示为

$$w_{ji}^{d}=\frac{\boldsymbol{F}_j^v\cdot\boldsymbol{T}_i^v}{\|\boldsymbol{F}_j^v\|\cdot\|\boldsymbol{T}_i^v\|} \tag{4}$$

其中，$\boldsymbol{F}_j^v$ 表示知识领域 j 特征向量；$\boldsymbol{T}_i^v$ 是文档 i 的特征向量，是由特征词与特征词在文档中的词频构成.

3.2 更新专家知识模型

在更新专家知识模型时要考虑专家上传文档的数量. 上传的数量多，说明该专家经验知识越丰富. 知识是有时效性的，所以要考虑时间因素，时间因素也反映了专家最近是否上传过文档[6].

专家 j 的知识模型通过专家知识向量 $\boldsymbol{E}_j^v$ 表示，由知识领域与在知识领域的知识水平构成：

$$\boldsymbol{E}_j^v=\{(k_1,w_{j1}^e),(k_2,w_{j2}^e),\cdots,(k_N,w_{jN}^e)\}$$

其中，w_{ji}^e 表示专家 j 在知识领域 i 的知识水平；e 表示 w_{ji}^e 是专家知识模型特征向量的权重.

$$w_{ji}^{e}=\Big(\sum_{p\in D_j^s}((1+w_{pi}^d)^{T_p}-1)\Big)^{T_j} \tag{5}$$

$$T_j=\frac{N(D_j^s)}{\sum_{p\in D_j^s}T_p} \tag{6}$$

其中，T_p 表示时间因素，是与具体文档相关的. 若文档 p 时效性高并且是最近上传的，那么 T_p 取较大的值；T_j 表示专家 j 上传的所有文档时间权重的均值；D_j^s 是专家 j 上传文档的集合；$N(D_j^s)$ 表示 D_j^s 中文档的数量.

4 知识需求建模

进行专家推荐的另一个重要步骤是用户知识需求的建模，只有准确建立用户的知识需求模型，才能推荐满足用户知识需求的专家.

知识需求包括直接知识需求与潜在知识需求. 直接知识需求表示用户明确知道其知识需求领域，已经阅读了文档，但仍然满足不了其知识需求，需要咨询专家. 潜在知识需求指用户无法明确具体知识需求领域，其浏览的文档所属知识领域分散，多发生在新用户或开始新的项目时.

4.1 直接知识需求建模

用户 U_j 的直接知识需求模型通过直接知识需求向量 $\boldsymbol{U}_j^{ev}$ 表示，是由知识领域与用户在知识领域的直接知识需求权重构成：

$$\boldsymbol{U}_j^{ev}=\{(k_1,w_{j1}^{en}),(k_2,w_{j2}^{en}),\cdots,(k_N,w_{jN}^{en})\}$$

用户 U_j 在知识领域 i 的直接知识需求权重 w_{ji}^{en} 通过式(7)进行计算：

$$w_{ji}^{en}=\frac{\Big(\sum_{p\in R_j^s}((1+w_{pi}^d)^{T_{jp}}-1)\Big)^{T_j}}{\sum_{f=1}^{N}\Big(\sum_{p\in R_j^s}((1+w_{pf}^d)^{T_{jp}}-1)\Big)^{T_j}} \tag{7}$$

$$T_j=\frac{N(R_j^s)}{\sum_{p\in R_j^s}T_{jp}} \tag{8}$$

其中，T_{jp} 表示用户 U_j 浏览文档 p 的时间因素，若文档 p 时效性高并且是用户 U_j 最近阅读的，那么 T_{jp} 取较大的值；T_j 表示用户 U_j 浏览文档时间权重的均值；R_j^s 是用户 U_j 浏览的文档的集合.

4.2 基于信息熵的潜在知识需求建模

用户 U_j 的潜在知识需求模型通过潜在知识需求向量 $\boldsymbol{U}_j^{iv}$ 表示，是由知识领域与用户在知识领域的潜在知识需求权重构成：

$$\boldsymbol{U}_j^{iv}=\{(k_1,w_{j1}^{in}),(k_2,w_{j2}^{in}),\cdots,(k_N,w_{jN}^{in})\}$$

在获得用户直接需求的基础上，计算用户 U_j 在知识领域 i 的潜在知识需求 w_{ji}^{in}，w_{ji}^{in} 通过下式计算：

$$w_{ji}^{in}=\mathrm{entropy}(k_i)=\mathrm{entropy}(k_{i1},\cdots,k_{im})=-w_{j,i1}^{en}\log w_{j,i1}^{en}-w_{j,i2}^{en}\log w_{j,i2}^{en},\cdots,-w_{j,im}^{en}\log w_{j,im}^{en} \tag{9}$$

其中,k_i 是知识领域 i;k_{im} 是与知识领域 i 直接相关的子级知识领域 im. 在领域本体中,叶节点的信息熵固定为 0.

4.3　用户知识需求模型更新

用户的知识需求模型如下式所示:

$$\boldsymbol{U}_j^v=\{(k_1,w_{j1}^u),(k_2,w_{j2}^u),\cdots,(k_N,w_{jN}^u)\}$$

其中,N 是知识领域的数量;u 表示 w_{jN}^u 是用户知识需求向量的权重. 用户 U_j 在知识领域 N 的知识需求权重 w_{jN}^u 由直接知识需求权重与潜在知识需求权重构成,如下所示:

$$w_{jN}^u=\alpha w_{jN}^{en}+(1-\alpha)w_{jN}^{in}\quad 0\leqslant\alpha\leqslant 1 \tag{10}$$

其中,α 是为了调节直接需求与间接需求的重要性. 为了避免噪音的干扰,对于通过式(10)获得的权重低于阈值 δ 的知识领域将被删除.

5　多知识领域的专家推荐

多知识领域的专家推荐包括两个步骤:计算专家知识模型与用户知识需求模型的匹配度,然后将匹配度最高的专家推荐给用户.

5.1　专家知识模型与用户知识需求模型匹配

专家 E_e 与用户 U_u 匹配度采用内积计算:

$$f(e,u)=\boldsymbol{E}_e^v\cdot\boldsymbol{U}_u^v \tag{11}$$

其中,$\boldsymbol{E}_e^v$ 是专家 E_e 的知识模型;$\boldsymbol{U}_u^v$ 是用户 U_u 的知识需求模型.

5.2　专家推荐

将匹配度最高的专家推荐给用户.

$$E_u=\{e^*\mid f(e^*,u)=\max_{e\in E}f(e,u),e^*\in E\} \tag{12}$$

其中,E_u 是推荐给用户 u 的专家集合;E 是所有专家集合.

6　应用实例

为了验证专家推荐方法的有效性,改进了具有专家推荐功能的知识管理系统. 该系统已经被应用在某航空部门中. 在系统中的专家黄页中,专家可以属于多个知识领域,只是在这些知识领域中的知识水平各有不同,这样可以全面的对专家知识能力进行描述,用户可以通过多个知识领域在专家黄页中寻找到其所需要的专家,在进行专家推荐时采用本文提出的基于模糊文本分类的多知识领域专家推荐方法,图 3 为新系统截图.



图 3　新系统截图

7　结 束 语

本文提出了一种多知识领域专家推荐方法,建立了模糊文本分类器,在此基础上构建了专家知识模型与用户知识需求模型,给出了多知识领域的专家推荐算法. 提高了专家推荐的准确性,也提高了知识共享的效率. 通过在航空知识管理系统中的成功实施验证了本方法的可行性与有效性.

参考文献(References)

[1] Maybury M T. Expert finding systems [R]. MTR06B000040, 2006

[2] Yang K W, Hu S Y. Automatic expert identification using a text categorization technique in knowledge management systems [J]. Expert Systems with Applications, 2008, 34(2): 1445－1455

[3] Fang H, Zhai C. Probabilistic models for expert finding [C]//Amati G, Carpineto C, Romano G. Proceedings of the 29th European Conference on Information Retrieval. Berlin: Springer-Verlag, 2007: 418－430

[4] Serdyukov P, Hiemstra D. Modeling documents as mixtures of persons for expert finding [C]//Macdonald C, Ounis I, Plachouras V, et al. Advances in Information Retrieval. Berlin: Springer-Verlag, 2008: 309－320

[5] Balog K, Azzopardi L, Rijke M D A. Language modeling framework for expert finding [J]. Information Processing & Management, 2009, 45(1): 1－19

[6] Hughes A M. The complete database marketer: Second generation strategies and techniques for tapping the power of your customer database [M]. New York: McGraw-Hill, 1996

(编　辑:娄　嘉)