一种有效的基于本体的词语-概念映射方法

李 文 陈叶旺 彭 鑫 赵文耘

(复旦大学计算机科学技术学院 上海 200433)

摘 要 词语-概念映射是基于本体的语义检索的重要一环,对语义检索的查准率及查全率有很大的影响。在传统的基于关键词匹配的方法中,通常从词语-概念的共现程度来计算它们的相关度,这种方法没有考虑概念的属性及属性值,即丢失了概念的语义信息。针对这一问题,提出了一种词语-概念映射方法,该方法基于本体三元组-文档标注结果,利用概念-文档与词语-文档两重关系,首先计算出词语-概念的相关度与置信度,再实现词语-概念的映射。实验结果表明,该方法有效地提高了检索的效果。

关键词 本体,词语,映射,标注,三元组

Effective Term-Concept Mapping Method Based on Ontology

LI Wen CHEN Ye-wang PENG Xin ZHAO Wen-yun

(School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract Term-concept mapping, as one of the key steps of ontology-based semantic search, has a great impact on precision and recall. In traditional methods based on keyword matching, term-concept co-occurrence is introduced to compute the association, but this kind of method doesn't consider the attributes of concepts or their values, losing the semantic information. To handle the problem, we proposed a term-concept mapping method based on triple-document annotation, which takes into count both the concept-document and term-document relationship, calculates the correlation and confident degree of term-concept firstly, and then gives the final mapping results. The experiments show that this method improves the precision and recall effectively.

Keywords Ontology, Term, Mapping, Annotation, Triple

1 引言

基于语义的查询,作为从20世纪90年代问世的语义网[3]在信息检索领域的一个应用,已经在改进检索查全率和查准率方面表现出巨大的潜力。与传统的基于关键词及词频的信息检索引擎相比,语义检索引擎在文本中加入语义tag,这种通过语义tag结构化、概念化文档中的对象的方法,比传统基于关键词搜索的方法在挖掘用户给出的信息中隐含的信息方面更加有效[1]。

在基于本体[3]的语义检索系统中,通常使用本体作为描述领域知识的基础,将文档标注到相关的本体知识概念中。在检索时,通过将关键词匹配到本体知识概念,并根据标注的信息找到相关文档。然而对于所有信息检索系统来说,存在着一个根本问题:关键词错误匹配[12],即系统的开发人员与用户使用不同的术语来描述相同的概念,如单车与自行车描述的是同一事物。这本质上是由于人类对世界的不同理解所造成的,同时也是不可避免的,但这对系统的检索结果有着很大的影响。此外,从用户对信息检索系统的使用情况来看,查询语句变得越来越简短,也使得词语错误匹配的问题变得更加突出。因此,无论是传统基于关键词匹配或者是基于语义的查询,在接受用户以自然语言(如关键词、疑问句)方式提交的查询之后,通常都使用查询扩展来提高查全率。这是因为通过查询扩展[2]可以利用与查询关键词相关的词语对查询进行修正,以找到更多、更准确的相关文档。

而在查询扩展中,一个关键的问题是如何找到与查询关键词相匹配的概念集。在基于本体的检索系统中,这个问题就转化为词语-本体知识之间的映射。词语-本体知识映射在本体和语义网领域中有着大量的研究,例如文献[16]使用相似度计算,在两个机器系统之间自动完成本体概念的映射。总的来说,现有的方法基本上是基于关键词方式来完成词语-知识映射的,这种方法常常会带来许多语义理解错误,文献[15]中称其为词语问题(vocabulary problems)。如同义词问题(synonyms)、歧义问题(polysemy)、异体问题(lemmas)、准同义问题(quasi-synonyms)等,在提高查全率的同时难以保证查准率。产生词语问题的根本原因在于,人们在现实生活中描述同样的对象或事件的用词存在着多样性。在基于本体的语义检索系统中,目前大多数利用文档对本体标注都是以单个本体知识-单个文档的形式进行的,虽然当中可以有多对

到稿日期:2009-11-18 返修日期:2010-01-20 本文受国家863高技术研究发展计划基金项目(2007AA01Z179)资助。

李 文(1985-),男,硕士生,主要研究方向为语义网,E-mail:072021123@fudan.edu.cn;陈叶旺(1978-),男,博士生,主要研究方向为语义网、本体论等;彭 鑫(1979-),男,讲师,主要研究方向为构件技术与软件体系结构、软件产品线、软件维护与再工程等;赵文耘(1964-),男,教授,博士生导师,CCF高级会员,主要研究方向为软件工程、电子商务及企业应用集成(EAI)。

多的关系，但我们认为这种方法丢掉了语义的信息。因此，在我们的方法中，考虑将文档标注以文档-本体知识关系来进行。本文提出了一种基于词语-概念三元组-文档标注结合的方法来计算词语-本体知识关系相关度的方法。该方法扩展了传统的单个词语到单个概念的方法，通过将词语（集）映射到本体中的知识关系，考虑上下文信息，以计算词语与三元组中对应知识的相关度。

本文第2节介绍已有的相关工作；第3节概述三元组（链）-文档标注方法；第4节介绍基于三元组-文档标注的词语-本体知识关系相关度计算；第5节描述实验、结果及与已有传统方法的对比；最后是总结。

2 相关工作

随着语义网的大幅进展，以及人们所面临的下一代搜索的挑战和语义万维网的兴起等，极大地推动了语义搜索这一潜在研究领域在最近的发展，并且促成了语义搜索这一名词的广泛使用。最近几年，明确提出或提到语义搜索的研究工作越来越多[4-7]。这些工作从不同领域的不同角度出发，以各种方式利用各种不同类型的语义信息来提高搜索的效果。

目前已经有了一些关于语义标注的研究工作。早期，一些研究者应用信息抽取技术对网页进行语义标注，例如Amilcare[4]工具的基于领域训练文档的标注方法；又如Ont-0-Mat[5]工具可支持对动态生成的网站的标注；结合以上工具和方法，S-Cream将Amilcare产生的标注结果与Ont-0-Mat定义的关系元数据内的概念标记匹配，但是由于Ont-0-Mat与Amilcare模型有较大的差别，S-Cream的效果并不好。类似匹配差异的问题在基于信息抽取技术的语义标注方法中常常出现，很难得到较好的解决。信息抽取技术获得的信息与本体中的定义（如类、实例、属性）无法很好地对应或匹配，导致信息的缺失，而使用额外地从抽取出的信息到本体中概念的匹配的方案又增加了额外的负担及信息丢失的风险。基于熵分类器语义角色标注[13]是浅层语义分析的一种可行方案，它描述了一个采用最大熵分类器的语义角色标注系统，把句法成分作为语义标注的基本单元，用最大熵分类器对句子中谓词的语义角色同时进行识别和分类，使用了一些有用的特征及其组合。在后处理阶段，在具有嵌套关系的结果中，只保留概率最高的语义角色。在预测了全部能够在句法分析树中找到匹配成分的角色以后，采用简单的后处理规则去识别那些找不到匹配成分的角色。以上几种方法可以归结为应用信息抽取技术的方法（Information Extraction，IE）。

此外，一些学者通过使用基于本体的信息抽取方法（Ontology Based Information extraction，OBIE）来实现标注。如SHOE[25]，它通过提供用户图形化的界面，帮助完成网页到SHOE本体的标注，此系统的缺点在于仅支持对静态网页的标注。其他如KIM[11]，在对文本标注的同时，将标注过程中发现的新概念补充进本体知识中。此类方法的普遍问题是仅关注本体中的实例本身与待标注文本之间的关系，而忽略了本体中实例与实例、实例与属性之间的关联，即语义信息。文献[21]则提出了一种有效的服务资源自动语义标注方法，该方法将服务语义标注过程分解为域标注和概念标注两个阶段，重点针对域标注问题提出了基于K-NN的域标注算法，并通过对实际服务资源的域标注实验，验证了该算法的有效性。

另外，基于自然语言处理的方法（Natural Language Processing，NLP）也是文档标注方法的一个重要分支，如Artquakt[6]和iOka[7]。上述两个方法考虑了谓语动词在句子中的重要性：它们常常可以对应到本体中的某些类的通用属性。因此借助通用语言本体来辅助实现从谓语动词到本体属性的映射。而RelEx[8]利用统计模型，计算词语出现的频度及其在语句中的语义位置，以获取表达属性的动词。同时结合句法分析将本体属性对应到上述标注的动词，但是该方法过高地估计了动词在语句中的重要程度，并没有考虑语句中由其他类型的词语对应本体属性的可能性。上述3个方法充分表明语句的句法结构对于标注有着相当重要的作用。通常来说，一个有着良好结构的句子，其各个成分与本体中实例及其属性值有很大的对应关系。此类方法中的共同难点是如何将谓语动词对应到领域本体中合适的类或实例的属性。为解决该问题，通常的做法是借助外部知识（如上述方法中的语言本体）来辅助完成这一映射。

对于基于本体的语义搜索，词语到本体知识的映射是搜索关键步骤之一，匹配成功与否直接影响到最终的查准率和查全率。不同用户之间、用户与开发者之间对相同概念的理解偏差和用户无法准确地提供所需的信息，是影响词语-本体知识关系匹配的两个重要因素。计算相关度的常用方法有余弦相似度、Dice相似度。词语共现性是计算相关度的重要依据。通常来说，经常同时出现的词语相关度一般比较大。而计算共现性时，比较通用的依据有词语粒度、短语粒度[12]、概念粒度[24,25]等等。由Xu和Croft提出的LCA[12]方法是一种经典的语义查询扩展方法，它最主要的贡献在于通过计算初始检索结果集的top-k文档中词语和查询中词语的共现度，来找到查询相关词语作为扩展词语。此外，计算相关度时还有一些其他重要因素，如句法上下文[20]、信息熵[21]等。文献[10]中提出了一个语义构件检索的理想场景，即通过语义推理和一系列问答式的会话过程，将一个本体中的概念映射到另一个本体中的概念。而文献[22]中提出一种通过用户与构件库系统之间的会话式过程完成这一本体协商任务的方法，该方法使用动宾短语中谓语和宾语的联合匹配，结合状语及定语对在谓语和宾语刻面属于词汇集合的概率进行映射。

3 三元组-文档标注方法

目前的语义标注工作，大部分还都只是用单个本体的Individual（即实例）为文档作标注，通过在文档中查找是否存在与Individual相关的词汇来实现。我们认为这种方式割裂了Individual所存在的语义环境，使得资源标注的结果不够准确，因为一个概念并不是单独存在的，它必须与其相关的各种属性及属性值共存，才能体现出其语义，否则与普通的文本就没有区别了[22]。我们将用单个本体的individual对文档进行标注的方法的不足归纳为以下3点：

（1）有些词汇意思相同，却常有不同的表达词组。如果只用Individual标签值来做统计，往往会造成统计结果不准确（缺失了Individual的各个属性值）。

（2）在含有英文的文本值中，单词的单复数形式、动词的过去式、动词完成式等也常使统计结果不准确。

（3）最重要的是，这种方式割离了Individual的存在语义环境，会使标注的结果可能完全错误。

由此我们提出，语义标注方式应该是基于本体语义关系的，即本体三元组或三元组链的形式，亦即相当于用本体语义图中的子图来为资源作标注，不应当只是通过某个单独的本体知识点来进行标注。

因而，我们之前的工作[14]采用非内嵌方式来标注文档。从不同角度上来看，我们的方法属于半自动化方法。标注的结果存储在数据库中，而不是创建一个标注本体。每条标注都至少包含3方面内容：一个是本体知识关系，以三元组表示；第二个就是资源，包括资源标识和地址；第三个是前二者之间的相关度。同时，一个本体知识关系可以标注多个资源，一个资源可以被多个关系所标注，即二者之间可以是多对多的关系。表1为一个标注示例。

表1 语义标注结果

知识关系	资源	相关度
〈黄瓜，是，植物〉	www.fao.org/2009003/173.rdf	0.81
〈白斑病，危害，黄瓜〉	www.fao.org/2009003/173.rdf	0.65
〈百菌清，治，白斑病〉	202.120.224.175/med/163.doc	0.90
…		

为了方便叙述，我们进行以下假设约定：设 RB 是一个信息资源库，包括各类文档，记为 $DS=\{d_1,d_2,\cdots,d_i,\cdots,d_m\}$，其中 d_i 是第 i 个文档，m 为文档数量，$1\leqslant i\leqslant m$。$T=\langle subj,pred,obj\rangle$ 表示一个三元组。设 TS 是用来标注资源的本体知识集，记为 $TS=\{t_1,t_2,\cdots,t_j,\cdots,t_k\}$，其中 t_j 是第 j 个三元组，k 为实例数量，$1\leqslant j\leqslant k$。若一个文档 d 受本体知识关系 t 标注，则记为 $\succ_d^t=\langle t,d,r\rangle$，其中 t 表示一个只是关系，d 表示一个文档资源，r 是前二者之间的相关度。

定义1(语义标注) 语义标注是从知识库文档库到标注结果的映射，记为 $\delta: DS\times TS\rightarrow\{\succ_{d_i}^{t_m}\}$，其中 $0<i\leqslant|DS|$，$0<m\leqslant|TS|$。

4 映射方法

4.1 词语-概念相关度

在词语-标注文档-本体知识的相关关系中，词语可能被包含在多个文档中，而每个文档又可被一个或多个本体知识三元组标注。通过统计包含词语的文档所属的本体三元组，可以统计出这个词语对不同三元组的相关程度，这种相关程度说明了词语-本体知识三元组间的联系紧密程度。

为计算这种相关关系，我们通过以下规则来说明自然语言词汇对本体知识的相关关系。

规则1 如果词语集合通过文档映射到的三元组越多，那么其对单个三元组的所属程度越低。

规则1是从三元组空间的分布来分析。根据直觉，如果词语集合与越多的三元组相关联，说明词语集合对于这些三元组的区分度就越低，与三元组的关联度也低。

规则2 如果词语集合通过越多的文档映射到同一个三元组，那么词语集合对此三元组的所属程度越高。

规则2是从三元组的文档空间来分析。如果词语集合在用三元组标注过的越多文档中出现，说明这些词语在文档中分布越广越普遍，因此其与三元组的关联度也越高。

规则3 若词语集合通过文档映射到某个三元组。如果三元组中对应本体实体标签词汇在词语集合中出现的个数越多，那么词语集合与三元组的所属程度越高。

规则3是从三元组对应本体片段的标签词汇空间来分析的，对于一个三元组(有3个组成部分：[主(subj)、谓(pred)、宾(obj)])，当3个组成部分的标签词汇都出现在词语集合中时，三元组与词语集合的相关度最高；当有2个组成部分的标签词汇出现在词语结合中时，相关度次之；只有1个组成部分的标签词汇出现在词语集合中时，相关度最低。

基于以上规则，我们给出计算三元组 t 与词语集 c 的关联程度的公式如下：

$$w(t,c)=\frac{tf(t,c)\times\log(n_t/N)}{\sqrt{\sum_{t\in c}[tf(t,c)\times\log(n_t/N)]^2}} \tag{1}$$

式中，$w(t,c)$为 t 与 c 的关联程度，而 $tf(t,c)$为 t 在词语集 c 中出现的次数，N 为 t 所标注过的词语集总数，n_t 为词语集中出现 t 的文本数，分母为归一化因子。

规则4 令出现过词语集 c 的文档集合为 ds_c。而知识关系三元组 t 标注的文档集合为 ds_t。二者交集越大，置信度越高；二者并集越大，置信度越低。

规则4是同时考虑词语集出现的文档集合与三元组标注的文档集合。如果词语集出现过的文档集与知识关系三元组标注过的文档集的重合程度越大，那么词语集与三元组的置信度也越高。

由此，我们给出词语集与三元组的置信度：

$$\tau(c,t)=\frac{ds_c\bigcap ds_t}{ds_c\bigcup ds_t} \tag{2}$$

式中，ds_c 表示出现过词语集中任一词语的文档集合，ds_t 表示三元组标注的文档集合。

由此，我们给出算法1，计算考虑词语集合关联到三元组的文档空间以及三元组本身标注的文档空间之下词语集合与三元组的置信度。于是，我们得到词语集 c 与三元组 t 的相关度：

$$r(c,t)=w(c,t)\times\tau(c,t) \tag{3}$$

Algorithm 1

输入数据：与词语集合 d 相关的三元组及相应所属相关度二元组的集合 Γ

输出数据：与词语集合 d 相关的三元组及相应置信度二元组的集合 Π

0 for each $w(t,c)$ in Γ
1 for each *word* in c
2 for each *doc* in DS
3 如果 *word* 在 *doc* 中的词频大于阈值且 $doc\notin ds_c$
4 那么 把 *doc* 加入到 ds_c 中
5 如果 *doc* 受 t 标注过
6 那么 把 *doc* 加入到 ds_t 中
7 计算 $\tau(c,t)=\frac{ds_w\bigcap ds_t}{ds_w\bigcup ds_t}$
8 将 τ 加入到 Π
9 end

4.2 查询-文档相关度

计算查询词语与本体中概念相关度的目的是通过词语与概念的关系获取标注在本体概念之上的文档。下面给出查询-文档相关度的计算：

$$r(c,d)=r(c,t)\times r(t,d) \tag{4}$$

式中，$r(c,d)$是词语集与本体知识三元组的相关度，$r(t,d)$是对应本体三元组与文档的相关度。本方法扩展了课题组之前的工作中的语义标注方法[14]，原方法中考虑了文本上下文到本体的对应关系，而本方法基于文档-本体知识的标注，考虑

了词语(集)-文档以及文档-本体知识的双重因素，将词语(集)映射到本体中相关的知识片段。

5 实验及比较

5.1 实验设置

目前在语义搜索领域还没有一个公认的测试数据集和评价方法，因而我们在做测试时，使用的数据是由我们自己根据国际粮农组织[9](FAO)建立并公布的本体、抽取其中的部分建立的本体知识，以及经过我们的语义标注工具或手工方式标注过的农业相关的资源。本体知识使用的是农业病虫害本体 CropDisease、花卉知识本体 Flower 以及足球本体 Soccer，其相关信息如表 2 所列。对应资源分别是来自中国农科院作物品种资源研究所依据《中国粮食作物、经济作物、药用植物病虫原色图鉴》、《中国农业百科全书》制作的农作物病虫害知识[19]、上海花卉网的资源[18]以及新浪足球新闻网国际足球新闻，相关信息如表 3 所列，以上标注结果由本课题组在之前的研究中给出[14]。沿袭信息检索领域的传统，使用 *Precision@n* 和 *Recall@n* 作为主要的评价方法和指标，其中 *Precision@n* 表示前 n 个结果中的查准率，*Recall@n* 表示前 n 个结果中的查全率。其计算公式如下：

$$Precision@n=\frac{\#\ \text{of true result in top-}n\ \text{results}}{n} \tag{5}$$

$$Recall@n=\frac{\alpha\cap\beta}{\beta} \tag{6}$$

式中，α=检索的前 n 个结果文档中实际相关的文档集合，β=人工判断的检索结果中前 n 个结果。

表 2 实验本体统计信息

本体名	概念数	实例数
Flower	113	2,400
Soccer	54	1230
CropDisease	274	3,730

表 3 文档集的统计数据

文档集	文档总数	平均词数	实例平均标注文档数
花卉知识	119	≈987	≈3.6
新浪国际足球新闻	403	≈1135	≈5.4
农作物病虫害知识	1,119	≈1129	≈4.3

同时，我们选择未经优化的原始概念检索方法作为基本对比方法(RawMethod)。由于大多数用户在检索过程中只考虑前 5～20 个结果，因此我们取 n=20 分别进行试验。在实验过程中，共构建了 100 个查询，对每个查询返回结果的前 20 个文档的相关性进行统计分析。

5.2 实验结果及分析

表 4 给出了两种方法在 3 个文档集上的实验结果评价。

表 4 3 个文档集上的实验结果

衡量标准	文档集	基本方法	改进方法
Precision@20	花卉知识	0.285	0.330
	新浪国际足球新闻	0.280	0.322
	农作物病虫害知识	0.283	0.338
Recall@20	花卉知识	0.202	0.254
	新浪国际足球新闻	0.185	0.240
	农作物病虫害知识	0.200	0.250

图 1 和图 2 给出了不同方法在 3 个文档集中的检索结果。

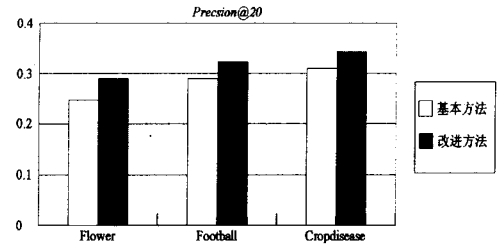

图 1 两个方法在不同文档集上的 Precision@20 的比较

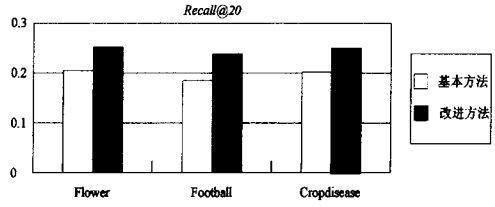

图 2 两个方法在不同文档集 Recall@20 的比较

从实验结果来看，首先考虑本体规模对结果的影响。当本体的规模增大时，检索结果的查准率及查全率并没有太大的改变，由此可以看出本体规模对于检索效果影响不大。其次，从文档集的数量来衡量，农作物病虫害知识的文档集最大，花卉知识的文档集最小，相差约 10 倍，但是检索效果相差小于 2%，由此可以得出文档集的大小对于检索效果几乎没有影响。再对比原始概念检索方法，我们的方法在查准率及查全率上都有一定的提高，主要原因是方法考虑到本体中的上下文语义信息，而不是简单地统计本体知识标签出现的概率。因此，本文方法对于查准率以及查全率的提高有一定的效果。

结束语 随着语义网的出现，其在信息检索领域的应用——基于语义的检索相对传统的基于关键词检索是一个显著的进步，因为它能较好地解决基于关键词很难处理的词语问题。而其中词语-概念映射也是对于检索结果起着重要影响的一环。本文提出了一种词语-概念映射方法，利用概念-文档与词语-文档两重关系，计算出词语与概念的相关度与置信度，再实现词语(集)到本体概念(集)的映射。从实验结果来看，有效地提高了检索的效果。

参 考 文 献

[1] Dong Hai, Hussain F K, Chang E. A Survey in Semantic Search Technologies[C]// 2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST). 2008: 403-408

[2] 田萱，杜小勇，李海华. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报，2008，19(8)：2043-2053

[3] www. w3. org/2001/sw/WebOnt/ WebOntology

[4] Ciravegna F, Wilks Y. Designing adaptive informationextraction for the Semantic Web in amilcare[G]. Handschuh S, Staab S. eds. Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications, Amsterdam: IOS Press, 2003: 112-127

[5] Handsehuh S, Staab S, Maedche A. CREAM: Creating relational metadata with a component-based, ontology-driven annotation framework [C]// Proc. of the 1st Int'l Conf on Knowledge Capture. New York: ACM, 2001: 76-83

[6] Alani H, Kim S, Millard D, et al. Automatic ontology-based

knowledge extraction from Web documents[J]. Intelligent Systems,2003,18(1):14-21

[7] Lai Y,Wang R. Towards automatic knowledge acquisition from text based on ontology—centric knowledge representation and acquisition[C]// Proc. of the K—CAP 2003 Workshop on Knowledge Markup and Semantic Annotation (Semannot' 2003). 2003

[8] Schutz A, Buitelaar P. RelExt: A tool for relation extraction from text in ontology extension[C]//Proc. of the 4th Int'l Semantic Web Conf(ISWC). Berlin:Springer,2005:593-606

[9] http://www.fao.org

[10] Christophides V,Karvounarakis G,Plexousakis D,et al. Optimizing taxonomic semantic Web queries using labeling schemes [J]. Journal of Web Sematics,2003,1(2):207-228

[11] Popov B, Kiryakov A, Ognyanoff D, et al. KIM: A Semantic Platform for Information Extaction and Retrieval[J]. Journal of Natural Language Engineering,2004,10(3/4):375-392

[12] Xu J X,Croft W B. Improving the effectiveness of information retrieval with local context analysis[J]. ACM Trans. on Information Systems,2000,18(1):79-112

[13] 刘挺,车万翔,李生. 基于最大熵分类器的语义角色标注[J]. 软件学报,2007,18(3):565-573

[14] Chen Ye-wang,Li Wen,Peng Xin,et al. An Improved Semantic Annotation Method for Documents Based on Ontology[C]// CSWS. 2009

[15] Furnas G W,Landauer T K,Gomez L M,et al. The vocabulary problem in Human-System communication[J]. Communications of the ACM,1987,30(11):964-971

[16] Salton G, McGill M. Introduction to Modern Information Retrieval[M]. New York:McGraw-Hill,1983

[17] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval [M]. New York:Addison-Wesley-Longman,1999

[18] www.flower-sh.cn/article_list.asp? c_id=74&page=4

[19] http://icgr.caas.net.cn/disease/

[20] Gao J, Zhou M, Nie J Y, et al. Resolving query translation ambiguity using a decaying Co-occurrence model and syntactic dependence relations[C]//Järvelin K,Chairs P,Baeza-Yates R,et al.,eds. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press,2002:183-190

[21] Jang M G,Myaeng S H,Park S Y. Using mutual information to resolve query translation ambiguities and query term weighting [C]//Dale R,Church K,eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. College Park: Association for Computational Linguistics,1999:223-229

[22] 蔡怡峰,彭鑫,钱乐秋. 面向语义构件检索的交互式查询方案生成[J]. 电子学报,2008,36(8):1631-1636

[23] 陈叶旺. 国家农业本体协同建构与语义检索若干技术研究[D]. 上海:复旦大学,2009

[24] Chang Y,Ounis I,Kim M. Query reformulation using automatically generated query concepts from a document space[J]. Information Processing and Management,2006:42

[25] Zhang M,Song R H,Ma S P. Document refinternet based on semantic query expansion[J]. Chinese Journal of Computers, 2004,27(10):1395-1401

(上接第 134 页)

(3)运用模糊理论方法估算软件开发成本,是用已完成的软件成本估算欲开发的软件成本,可以节省细化软件开发过程或软件组成的时间。有经验的估算人员通过已完成软件的开发成本,利用相似法,加上系数调整,就可以进行估算。受此启发,把相当数量的软件成本及软件特性进行整理输入计算机,并和人工智能技术相结合,就可实现比较准确的成本估算。

(4)模糊理论对软件成本影响因素层次性问题有极好的适用性,能将影响软件开发成本的各因素综合起来,能产生符合客观实际的结果。

软件开发成本估算与模糊理论方法具有很强的结合性,这种结合可以很好地进行软件开发成本估算,同时可以保证结果的客观性,对软件开发起到了积极的指导作用。

参 考 文 献

[1] Li J,Ruhe G,AlEmran A,et al. A Flexible Method for Effort Estimation by Analogy [J]. Empirical Software Engineering, 2006,12(1):65-106

[2] ISBSG. Estimating, benchmarking & research suite release 9 [DB]. Hawthorn, Australia: International software Benchmarking Standards Group—ISBSG,2005

[3] 王祯显,廖小建,杜晓玲. 工程造价快速估算新方法及其应用[M]. 北京:中国建筑工业出版社,1998

[4] 朱训生. 工程管理的模糊分析[M]. 上海:上海交通大学出版社,2004

[5] Boehm BW,Valerdi R,Lane J,et al. COCOMO suite methodology and evolution[J]. CrossTalk:The Journal ofDefense Software Engineering,2005,18(4):20-25

[6] Briand L C,Wieczorek I. Resource Estimation in Software Engineering[M]//Marcinak J J,ed. Encyclopedia of Software Engineering. New York:John Wiley & Sons,2002:1160-1196

[7] 任永昌,邢涛,于忠党,等. 数据库规模估算数学模型研究[J]. 微电子学与计算机,2009,26(7):36-39

[8] 李明树,何梅,杨达,等. 软件成本估算方法及应用[J]. 软件学报,2007,18(4):775-795

[9] Dewson R. Beginning SQL Server 2005 for Developers[M]. Beijing:Post & telecom press,2006

[10] Nguyen H T,Wang T H,Wu B L. On probabilistic methods in fuzzy theory[J]. International Journal of Intelligent Systems, 2004,19(1):78-80

[11] Ruhe M,Jeffery R,Wieczorek I. Cost estimation for web applications[C]//Proc. 25th Int'l Conf. Software Engineering. Los Alamitos,CA:IEEE Computer society Press,2003

[12] Jorgensen M. A review of studies on expert estimation of software development effort[J]. Journal of Systems and Software, 2004,70(1/2):123-128