doi:10.3969/j.issn.1001-0505.2009.06.000

基于本体的文档语义标注改进方法

陈叶旺 李 文 彭 鑫 赵文耘

(复旦大学计算机科学技术学院,上海 200433)

摘要:在领域本体知识的语义环境和资源文档结构基础上,提出一种文档语义标注改进方法,分析、计算标签—文档的词频相关性和语义环境在局部窗口的共现性,实现对各类文档资源的语义标注.该方法首先提取出文档资源的纯文本内容,并分解出子句、句和段落集合.然后,对于每个具体的领域知识项,在本体知识库中寻找其语义环境信息.最后,按照 7 条相关度规则,分别计算出这些信息与分解后文档内容的相关度,从而完成整个文档库内和知识库内的综合计算,得到该项知识与文档资源的最终相关度.实验结果显示,该方法能够依据领域本体,有效地对互联网中大量以网页等形式存在的多种类文档知识资源进行自动语义标注.

关键词:本体;语义环境;语义标注

中图分类号:TP301 **文献标志码**:A **文章编号**:1001-0505(2009)06-0000-00

Improved semantic annotation method for documents based on ontology

Chen Yewang Li Wen Peng Xin Zhao Wenyuan

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

Abstract: Based on the semantic context and the structural info of a document, an improved semantic annotation method is proposed. The correlation between the ontology entity and the document and the co-appearance of the label-words frequents and the semantic context in local window are analysed and calculated. Firstly, this method extracts the text content from the document, and then decomposes it into a sub-sentences set, a sentences set and a paragraphs set. For each knowledge item in ontology, the context information of the item is extrated, and then the correlation between these information and those decomposed documents sets is calculated. Finally, the final correlation between the knowledge item and the document in the range of all document base and ontology base are obtained. The experimental results show that based on domain ontology, this method can annotate unstructured documents in web automatically and effectively.

Key words: ontology; semantic context; semantic annotation

随着万维网的飞速发展,Web 上网络资源也越来越丰富.为了提供集成化的领域综合知识服务平台,需要在相关知识资源收集、整理的基础上提供一种高效智能的检索手段.然而,现有的资源大都以非结构化的文档形式存在,这些文档是为人类阅读准备的,不能面向机器处理.另一方面,语义网技术[1]通过本体规范表达领域知识,并使计算机识别和处理这些知识.针对这 2 个方面的差距,为资源提供语义标注变得紧迫且必须.

目前,针对海量数据的自动或半自动工具已相即出现[2-3],主要包括:① 基于传统信息抽取技术的方法.例如,工具 Amileare[2] 应用机器学习的方法在标注好的训练集上进行训练,通过提供不同领域的标注训练文档,适应多种领域的需求[3].② 基于本体信息抽取技术的方法.该方法将本体作为信息抽取过程中可用资源的一部分,利用本体内已有实例信息来构造列表,简化抽取过程中对于概念实例的识别.例如,方法 SemTag 在 TAP 本体实例集

收稿日期:2009-06-30. **作者简介**:陈叶旺(1978—),男,博士生;赵文耘(联系人),男,教授,博士生导师,ywzhao@fudan.edu.cn.
基金项目:国家高技术研究发展计划(863 计划)资助项目(2007AA01Z179).
引文格式:陈叶旺,李文,彭鑫,等.基于本体的文档语义标注改进方法[J].东南大学学报:自然科学版,2009,39(6): - . [doi:10.3969/j.issn.1001-0505.2009.06.000]

合中查找所有与待标注词匹配的可能实例集合,然后按照待标注词的上下文与实例集合中每个实例的上下文分别构造各自的文本向量,进行相似度计算,找到与待标注词最匹配的实例. ③ 基于自然语言处理的方法. 为了处理自然语言超文本数据,文献[3-5]都试图从句子的主谓宾语法成分中找到对应的 RDF 陈述.

虽然上述方法都能取得不错的效果,但存在着一些不足:① 通常只识别实体的词汇或标签,无法识别文本中表达的行为、动作和关系. ② 自然语言处理方法[3-5]的处理语句仅局限于对主谓宾关系的分析,未能有效应用于其他句法关系. ③ 在映射谓语动词到本体属性时,需要利用外部领域知识,但通用语言本体却不能完全涵盖领域内词汇. ④ 多数研究工作是面向英文的,中文文档语义标注的研究还不充分.

针对这些问题,本文提出了一种基于本体的文档语义标注改进方法. 利用领域本体所表达的丰富语义环境信息,通过分析领域本体知识的语义环境和资源文档结构 2 方面信息,实现对农业领域中各类文档资源的语义标注.

1 语义标注

简单而言,语义标注就是在本体中的知识点和资源之间建立关联,用以表示这些资源是知识点的一个扩展描述.

标注工作可按自动化程度划分为 2 种:完全的人工手动标注和由工具实现的半/全自动化标注. 人工标注的可信度固然高,但是面对海量资源,仅靠人工标注几乎是不可能完成的,因此自动标注工具也是必不可少的. 利用本体对资源的语义标注,需要如下 3 个基本要素:

1) 标注对象 即各种信息资源,包括各类有格式和无格式的文本文档,其形式化表达为

$$\mathrm{ds} = \{d_1, d_2, \cdots, d_i, \cdots, d_m\}$$

式中,ds 表示资源库,即一个信息资源库包括各类文档;d_i 表示资源库中第 i 个文档;m 表示文档数量,且 $0 < i < m$.

2) 标注知识 指用一个或多个领域本体描述的概念、实例或关系,而这些概念、实例或关系也同样是文档库文档中所描述的内容. 语义标注依赖于某个特定领域知识,而非普适知识. 领域本体知识成分可分为元知识和知识实例. 前者描述的是抽象领域知识关系,后者则表述具体知识的真实存在,因而适用于标注文档的知识源[6]. 本文中,知识库(kb)是指用于标注资源的本体知识实例,记为 $\mathrm{kb} = \{\mathrm{ind}_1, \mathrm{ind}_2, \cdots, \mathrm{ind}_k\}$,其中 k 表示实例数量.

3) 标注方式 按照标注保存形式可将标注方式划分为内嵌式和非内嵌式 2 种. 内嵌标注方式比较容易实现,但由于已经把标注作为本体内容的一部份,因而修改起来比较困难. 此外,内嵌标注方式把本体变得复杂,也增加了本体知识库维护的负担. 因此,选用非内嵌方式来标注文档,标注结果存储在数据库中,而不是创建一个标注本体. 若一个文档 d 受本体知识 ind 标注,则记为 $\succ_d^{\mathrm{ind}} = \langle \mathrm{ind}, d, r \rangle$,其中 r 表示二者之间的标注相关度.

定义 1(语义标注) 语义标注是指从知识库和文档库到标注结果的映射,记为 $\delta: \mathrm{ds} \times \mathrm{kb} \rightarrow \{\succ_{d_i}^{\mathrm{ind}_m}\}$,其中 $0 < i < |\mathrm{ds}|$, $0 < m < |\mathrm{kb}|$.

2 语义标注方法

2.1 方法

为便于说明,首先进行以下假设:subsent 表示一个子句;句子 sent = {subsent} 表示该句是一个子句集合;段落 para = {sent} 表示该段落是句子集合;文档 d = title + {para} 表示一篇文档由标题和段落集合组成.

文献[6]是根据知识实例的标签值在文档中出现的次数来标注文档的. 这种基本方法的计算规则如下:

规则 1 一个实例标签词语在一个文档中出现的词频越高,则与文档的相关程度越高.

根据这条规则可得

$$R(\mathrm{ind}, d) = \frac{\mathrm{count}(\mathrm{ind}, d) \times \mathrm{len}(\mathrm{ind})}{\mathrm{len}(d)} \tag{1}$$

式中,$R(\mathrm{ind}, d)$ 表示知识实例 ind 与文档 d 的相关度;$\mathrm{count}(\mathrm{ind}, d)$ 表示本体实例 ind 在文档 d 中出现的次数;$\mathrm{len}(\mathrm{ind})$ 表示 ind 标签的长度;$\mathrm{len}(d)$ 表示文档 d 的长度.

利用规则 1 进行标注的基本方法有许多不足. 例如,有些词汇具有同一含义,但常有不同的表达词组,如果只用individual标签值来统计,往往会造成统计结果不准确. 此外,这种方式割离了知识的存在语义环境,可能产生完全错误的标注结果.

为了更准确地实现语义标注,需要进行局部上下文分析和全局分析,主要途径是根据文档集中词语间共现性的统计数据来完成相关度计算. 这种方法认为语料库中经常共同出现的词语往往相关度很大. 分析共现性时,可以采用词语粒度、短语粒度[7]、概念粒度[8-9]等方法来实现. 然而,这些方法

大都利用文档或文档片断中包含的内容信息,忽略了从文档外部观察文档、实例、实例上下文之间的关系. 任何一个实体都不可以离开其所存在的环境而单独存在,而本体知识的语义环境应该体现在本体实体与其他实体的关系上.

定义 2(语义环境)　本体实体语义环境(context)是指实例在本体中与其他本体实体的关系集合,以本体三元组(triple)表示为 context(entity) = {triple | (triple. subject = entity) ∨ (triple. object = entity)}.

根据语义环境信息,提出了一个语义标注改进方法,该方法遵循以下规则:

规则 2　若一个实例标签词语在一个文档的标题中出现,则在其他条件相同的情况下,与这个文档的相关程度要比其他实例高.

规则 3　一个实例标注过的文档个数越多,与单个文档的相关程度越低. 一个文档被越多实例标注,则与单个实例相关程度越低.

规则 4　一个实例连同其某个属性值在一个文档中一起出现的词频越高,则与这个文档的相关程度越高.

规则 5　一个实例与越多的属性值出现在文档中,则与这个文档的相关程度越高.

规则 6　一个实例的对象属性值(即本体实例)与文档相关度越高,则与这个文档的相关程度越高.

规则 7　一个实例标签词语与其属性值在一个文档中物理距离越近,则与这个文档的相关程度越高.

在描述一个知识的文档中,这个知识和体现其存在的语义环境总是比较近的,体现在物理距离上(即文档中的文本距离)较短. 用段、句、子句将文档分为 3 个单位. 知识与其相关值在子句中一同出现的相关度最高,句子次之,段再次之.

基于上面几条规则,进行如下假设:

1) 令 I 为一个有限字符集;∇ 为自然语句子句分隔符集,且 $\nabla \in I$;ω 为自然语句句子分隔符集,且 $\omega \in I$;ζ 为自然语句段落分隔符集,且 $\zeta \in I$.

2) 令 subsent 为一个有限序列字符集,表示子句,其中 subsent. end $\in \nabla$,即其结尾字符为子句分隔符. 一个本体 triple 关系出现在文档 d 中的某个子句 subsent 中可记为 $\text{triple} \angle d_{\text{subsent}}$.

3) 令 $l^{\text{sub}}_{d,\text{triple}} = \{\text{subsent} \mid \text{triple} \angle d_{\text{subsent}}\}$ 表示在文档 d 中出现三元组 triple 的所有子句集合,则 $|l^{\text{sub}}_{d,\text{triple}}|$ 表示集合中的子句数量.

4) 令 sent 为一个有限序列 subsent 集,表示一个完整自然语言语句,其中 sent. end $\in \omega$,即其结尾字符为句分隔符. 一个本体 triple 关系出现在文档 d 中的句子 sent 中可记为 $\text{triple} \angle d_{\text{sent}}$.

5) 令 $l^{\text{sent}}_{d,\text{triple}} = \{\text{sent} \mid (\text{triple} \angle d_{\text{sent}}) \wedge (\neg\ \text{subsent} \in \text{sent}, \text{s. t. triple} \angle d_{\text{subsent}})\}$ 表示在文档 d 中出现 triple 的所有句子集合,则 $|l^{\text{sent}}_{d,\text{triple}}|$ 表示集合中的句子个数.

6) 令 para 为一个有限序列 sent 集,表示一个完整自然语言段落,其中 para. end $\in \zeta$. 一个本体 triple 出现在文档 d 段落 para 中可记为 $\text{triple} \angle d_{\text{para}}$.

7) 令 $l^{\text{para}}_{d,\text{triple}} = \{\text{para} \mid (\text{triple} \angle d_{\text{para}}) \wedge (\neg\ \text{sent} \in \text{para}, \text{s. t. triple} \angle d_{\text{sent}}) \wedge (\neg\ \text{sent} \in \text{para}, \text{s. t. triple} \angle d_{\text{sent}})\}$ 表示出现 triple 的所有段落集合,则 $|l^{\text{para}}_{d,\text{triple}}|$ 表示集合中的段落数量.

根据规则 3 ~ 7 规则可得

$$\Phi(\text{ind},d) = w \times \frac{\sum_{i=1}^{|\text{context}_{\text{ind}}|} \left(|l^{\text{sub}}_{d,\text{triple}_i}|^2 + |l^{\text{sent}}_{d,\text{triple}_i}|^{1.5} + |l^{\text{para}}_{d,\text{triple}_i}|^{1.2}\right)}{\log(10 + \text{len}(d))} \times \frac{|M|^2}{|\text{context}_{\text{ind}}|} \tag{2}$$

式中,$\Phi(\text{ind},d)$ 表示单个实例 ind 与文档 d 的内容相关度;$\text{context}_{\text{ind}}$ 表示实例 ind 的语义环境;$M = \{\text{triple} \mid (|l^{\text{sub}}_{d,\text{triple}}| > 0) \vee (|l^{\text{sent}}_{d,\text{triple}}| > 0) \vee (|l^{\text{para}}_{d,\text{triple}}| > 0)\}$ 表示所有在文档中子句、句子或段落中出现的本体知识三元组集合;$w = \log(N/n)$ 表示标注权重,其中 N 表示文档空间中所有文档的个数,n 表示用 ind 标注的文档的个数.

根据规则 2,可得到 d 与 ind 的相关度为

$$\text{prir}(\text{ind},d) = \begin{cases} 0 & \Phi(\text{ind},d) = 0 \\ (1 - w')\Phi(\text{ind},d) + w'\tau(\text{ind},d.\text{ title}) & \text{其他} \end{cases} \tag{3}$$

式中,$\tau(\text{ind},d.\text{ title})$ 表示布尔值,若实例 ind 的标签值在文档标题中出现则该值为 1,否则为 0;$w' = \frac{\text{len}(\text{ind})}{\text{len}(d.\text{ title}) \times \text{count\_ind}(d.\text{ title})}$ 表示权重,其中 $\text{len}(d.\text{ title})$ 表示文档 d 标题的文本长度,$\text{count\_ind}(d.\text{ title})$ 表示文档 d 的标题中出现本体实例标签的个数.

根据规则 6,考虑实例 ind 周边其他实例与文档的相似度,得到文档 d 与本体实例 ind 的综合相关度为

$$\text{pror}(\text{ind},d) = \begin{cases} \text{prir}(\text{ind},d) + \sum_{i=0}^{i=K} \dfrac{\text{prir}(\text{ind}'_i,d)}{\text{distance}(\text{ind},\text{ind}'_i)+1} & K>0 \\ \text{prir}(\text{ind},d) & \text{其他} \end{cases} \tag{4}$$

式中,$\text{prir}(\text{ind}'_i,d)$ 表示为 ind 的第 i 个对象属性值与文档 d 的原始相关度,其中 ind′ 表示 ind 的对象属性值;K 表示 ind 的对象属性个数;$\text{distance}(\text{ind},\text{ind}'_i)$ 表示 ind 与 ind'_i 之间的最近距离.

为比较方便,将综合相关度归一化为

$$\text{npror}(\text{ind},d) = \frac{\text{pror}(\text{ind},d) - \min(\text{pror})}{\max(\text{pror}) - \min(\text{pror})} \tag{5}$$

式中,$\text{npror}(\text{ind},d)$ 表示规一化后的相关度值.

2.2 语义标注算法

依据以上工作,列出相关度算法伪代码.

输入:本体知识库 ONP,文档库 ds.

输出:语义标注数据

for each doucement d in ds

分割文档 d,得到段落集 paras,句子集 sents,和子句集 subsents

for each individual ind in ONP do

计算以 ind 为 subject 或以 ind 为 object 的三元组集合 ts

for each triple in ts do

检查每一个 subs 中的子句,统计 $|l^{\text{sub}}_{d,\text{triple}}|$

检查每一个 sents 中的句子,统计得到 $|l^{\text{sent}}_{d,\text{triple}}|$

检查每一个 paras 中的段落,统计得到 $|l^{\text{para}}_{d,\text{triple}}|$

end

统计有出现在文档 d 中的三元组数量,即计算 $|M|$ 值

end

计算 $\Phi(\text{ind},d)$

if $\Phi(\text{ind},d)>0$ do

计算 $\tau(\text{ind},d.\text{title})$

依据式(4)和(5)计算 $\text{prir}(\text{ind},d)$ 和 $\text{nprir}(\text{ind},d)$

end;

end;

归一化处理所有相关度值,写入标注数据库.

下面对算法的复杂度进行分析. 在第 3 ~ 8 步循环中,给定一个本体知识个体和待标注文档,分别统计与本体个体相关的三元组在其子句、句子、段落中出现次数. 因为个体的三元组数量和文档长度都是有限的,因此,这一步循环的复杂度为常量 $O(C)$. 在第 2 ~ 10 步循环中,用本体知识库中所有的个体来进行统计工作,由于知识库内容会不断增加,个体数量也会随时间变化而增多,因而复杂度为 $O(n)$. 在第 0 ~ 16 步循环中,对文档库中所有文档进行标注运算,复杂度为 $O(n)$. 由此可知,整个算法的复杂度为 $O(n)\times O(n)=O(n^2)$.

3 实验与评测

为了考察本文方法的有效性和正确性,实验中建立 3 个不同规模的本体:足球本体、花卉知识本体和农作物病虫害领域本体. 本体的统计信息见表 1.

表 1 本体统计数据

本体	概念数	知识个体数
足球	54	1 230
花卉	113	2 400
农作物病虫害	274	3 730

对应于这 3 种不同的本体,采用 3 个检索文档集:新浪足球新闻网国际足球新闻、花卉知识文档(来自上海花卉网)和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. 将这些文档知识分别按不同格式(如 txt,html,xml,doc,pdf)进行转化,得到的文档集统计数据见表 2.

表 2 文档集的统计数据

文档库源	文档数	平均每个文档词汇数
花卉知识	119	987
新浪国际足球新闻	403	1 135
农作物病虫害知识	1 119	1 129

对这些测试数据集进行语义标注,每个有标注结果的实例都有相对应的被标注文档. 将这些文档进行人工统计、分析,并以人工处理结果作为比较基准,对方法的有效性进行评价,统计结果见表 3. 可以看出,结果符合规则 3,即采用基本方法时,一个实例标注的文档数要比改进方法中的相应数目高出很多,一个文档被标注的实例数也比改进方法中的相应数目高出很多. 这意味着基本方法标注结果的准确率比改进方法低. 此外,不同文档集的标注结果统计差异较大,且基本方法得到的结果的变化比改进方法更为剧烈,这也说明了改进方法的表现比较稳定. 统计结果在不同文档集上的差异可能是由于不同领域本体的详细程度不同所引起的. 例如,花卉测试本体比较简单,只描述一些基本简单关系,因而标注结果相对较少;足球新闻本体关

表 3 标注结果统计

文档库源	X		Y	
	基本方法	改进方法	基本方法	改进方法
花卉知识	17.4	3.6	0.86	0.18
新浪国际足球新闻	53.4	5.4	17.49	1.76
农作物病虫害知识	25.6	4.3	1.88	0.32

注:X 表示一个文档被标注的平均实例数量;Y 表示一个实例标注的平均文档数量.

系描述较细,标注结果也较多.

由于实验中有多个本体,为进行有效评测,可以在测试时对于不同本体选用与其对应的文档集进行标注. 根据文档规模,要从人力上判断文档集中所有文档与查询的相关性是难以完成的. 因此,采用以下 2 个标准进行评判:

1) 任选 a 个有标注结果的本体实例,取其前 k ($k=10$) 个标注结果,采用 precision[ind-doc]@(a,k)来衡量前 k 个被标注文档的准确率,计算公式为

$$\text{precision[ind-doc]@}(a,k) = \frac{1}{a}\sum_{i=0}^{i=a}\frac{l}{k} \qquad (6)$$

式中,l 表示受 ind 标注的前 k 个资源. 这样,将查准率和查全率结合,可以更全面地对前 n 个检索结果进行评价,这也符合大多数检索用户的习惯.

2) 任取 a($a=20$) 个被标注过的文档,采用 recall[doc-ind]@ {n,t} 来衡量它与文档相关度大于 t($t=0.40$) 的本体实例的查全率,计算公式为

$$\text{recall[doc-ind]@}(n,t) = \frac{1}{n}\sum_{i=0}^{i=n}\frac{p}{v} \qquad (7)$$

式中,p 和 v 分别表示采用自动方法和人工方法获得的与资源 doc_i 相关度大于 t 的本体实例个数.

方法的效率主要受以下几方面因素影响:①本体知识本身质量,包括知识表达方式和内容全面性;② 文档质量,包括文档内容文字表达、段落排版、有无错别字、文档格式等;③ 文档解析器质量. 图 1 为 2 种方法下查准率和查全率的实验结果. 可以看出,与基本方法相比,改进方法的查准率有大幅度提高,且查全率也接近基本方法. 结果表明,使用改进方法进行语义标注,可获得较高的查准率和查全率.

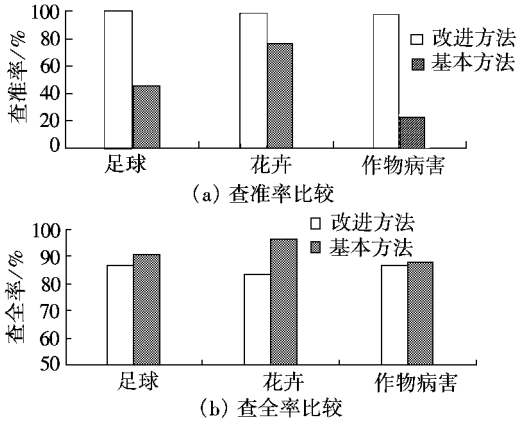

图 1　2 种方法的比较

4　结语

针对特定领域中大量以网页等非结构化形式存在的知识资源,提出了一种基于本体的文档语义标注改进方法,利用领域本体所表达的丰富语义环境信息,从标签—文档词频和语义环境在局部窗口的共现性 2 方面进行分析、计算,实现对各类文档资源的语义标注. 与已有的方法相比,该方法用来标注文档的是本体实例而非概念,且用来标注文档实例代表的是其背后的语义信息而非单独的实例. 实验结果表明,与以关键字方式进行标注的方法相比,该方法的准确率大幅提升,且查全率较为接近.

参考文献 (References)

[1] Berners Lee T, Hendler J, Lassila O. The semantic web [J]. *Scientific American Magazine*, 2001, **284**(5): 28 - 37.

[2] Ciravegna F, Wilks Y. Designing adaptive information extraction for the Semantic Web in amilcare [C]//*Annotation for the Semantic Web*, *Frontiers in Artificial Intelligence and Applications Amsterdam*. Amsterdam, Netherlands: IOS Press, 2003: 112 - 127.

[3] Alani H, Kim S, Millard D, et al. Automatic ontology-based knowledge extraction from Web documents [J]. *Intelligent Systems*, 2003, **18**(1): 14 - 21.

[4] Lai Y, Wang R. Towards automatic knowledge acquisition from text based on ontolog-centric knowledge representation and acquisition [C]//*Proceedings of the KCAP Workshop on Knowledge Markup and Semantic Annotation*. Sanibel, FL, USA, 2003: 111 - 127.

[5] Schutz A, Buitelaar P. RelExt: a tool for relation extraction from text in ontology extension [C]//*Proceedings of the 4th Internaltional Semantic Web Conf*. Berlin: Springer, 2005: 593 - 606.

[6] Vallet D, Fernández M, Castells P. An ontology-based information retrieval model [C]//*The 2nd European Semantic Web Conference*. Heraklion, Greece, 2005: 455 - 470.

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

[8] 张敏,宋睿华,马少平. 基于语义关系查询扩展的文档重构方法 [J]. 计算机学报, 2004, 27(10): 1395 - 1401.
Zhang Min, Song Ronghua, Ma Shaoping. Document refinement based on semantic query expansion [J]. *Chinese Journal of Computers*, 2004, **27**(10): 1395 - 1401. (in Chinese)

[9] Chang Y, Ounis I, Kim M. Query reformulation using automatically generated query concepts from a document space [J]. *Information Processing and Management*, 2006, **42**(2): 453 - 468.