基于语义互补推理的文献隐含知识的发现方法研究

温 浩[1] 温有奎[2]
(西安建筑科技大学信息与控制学院 西安 710055)[1] (西安电子科技大学经济管理学院 西安 710071)[2]

摘 要 文献知识发现已经成为解决海量信息检索难题的突破技术。但是目前的文献知识发现方法是基于词袋法的矢量空间模型方法。这类方法具有词汇元素之间语义无关性的先天不足,不能有效地发现文本之间存在的大量潜在知识。提出一种基于主谓宾(S,P,O)结构的最小知识单元表示及其语义推理的中文文献知识发现方法,避免了传统的文献知识发现方法的不足,并在此模型的基础上提出了一种推理算法,其能有效地发现文本中的潜在知识。经过实验证明,该方法与传统的文献知识发现方法相比有效地提高了潜在知识发现的正确率。

关键词 文本知识发现,语义互补推理,知识单元挖掘
中图法分类号 TP311 **文献标识码** A

Research on Discovery Method in Literature Implicit Knowledge Using Semantic Complementary Reasoning

WEN Hao[1] WEN You-kui[2]
(School of Information and Control Engineering, Xi'an University of Architecture Technology, Xi'an 710055, China)[1]
(School of Economics and Management, Xidian University, Xi'an 710071, China)[2]

Abstract The literature knowledge discoveries methods have become the breakthrough technologies to solve the difficult problem of massive information retrieval. But the current literature knowledge discovery methods are vector space methods based on word bag. This methods have congenital deficiency, i. e. semantic independence between vocabulary elements, so a large number of potential knowledge which exists between texts can't be discovered. Therefore, a new Chinese literature knowledge discovery method is put forward in this paper. The representation of the minimum knowledge unit and the semantic reasoning are based on subject predicate object (S,P,O) structure in this method. So the deficiency of traditional knowledge discovery method can be avoided, and a reasoning algorithm is proposed based on the structure to discover the potential knowledge between texts. The efficiency of discovering the potential of knowledge of our method is enhanced in contrast to the traditional methods, which is proved by the experiments in this paper.

Keywords Discovery in text, Semantic complementary reasoning, Knowledge element mining

1 引言

随着科学技术的飞速发展,科技文献的数量越来越多。虽然人们可以获得大量的科技文献,但是,科研人员因为存在背景知识的限制和思维方式的差异,能够从文献中获取的知识非常有限。因此,从大量的无序的文献中发现有用的知识,成为当今迫在眉睫的问题。D. R. Swanson 于 1986 年提出基于非相关文献的知识发现方法[1]。在大量的科技文献中,一些文献相互引用,文献间存在着显性联系,可以利用数据库引文检索到;另一些文献则互不引用或很少被共同引用,这类文献被称为是相互独立的,即非相关文献。一些非相关文献可通过各自的研究问题联系在一起,形成逻辑关联,称之为互补的非相关文献。互补的非相关文献之间的联系通过常规的数据库引文是检索不到的,是未被发现的隐含联系,而且隐含联系的数量可能远高于相互引证的显性联系的数量。通过分析处理非相关文献,能够获得大量的潜在知识,并且所能获得的知识量会远多于通过分析相关文献而获得的知识量。实验证明了通过分析不相关文献而获得的潜在知识具有更大的创造价值。在这种情况下,2000 年 Swanson 为此荣获 ASIST(American Society for Information Science and Technology)最高荣誉奖[2]。自 Swanson 之后,不断地有大量学者在 Swanson 方法的基础上进行改进,提出了更多的文献知识发现方法。1993 年,Z. Chen 在 Swanson 的方法的基础上,提出把知识发现深入到文献内部的知识片研究[3];2001 年,Weeber[4] 提出把以往的知识发现方法和语义分析方法进行结合;2006 年,Illhoi Yoo 提出引入语义文本挖掘方法[5];Tsuruoka[6] 和 Yetisgen[7] 提出将统计词汇相似性的研究方法提升为对文献中和文献之间词汇的共现统计(the co-occurrence statistics)的方法;2010 年 T. Cohn[8] 进行了概念之间高阶关联研究;2012 年周峰等[9] 在传统方法的基础上加入概念的语义相似度,寻找合理的假设发现。以往的研究方法基于单词、短语表达非相关文献概念,没有解决文献知识片段的语义关系问题,

到稿日期:2013-08-05 返修日期:2013-11-25
温 浩(1979—),男,博士后,主要研究方向为模式识别与智能系统,E-mail:smczg@126.com;温有奎(1951—),男,博士,教授,主要研究方向为智能推理,E-mail:wykui123@126.com。

使得非相关文献发现方法难以进入自动识别和推理研究。基于文献的知识发现技术越来越受到重视，2005 年，H. Roger 在文献[10]中指出，基于文献的知识发现技术在未来的 12～15 年后将从一种思想转化为市场医疗产品。产业部门的估计显示 90%的新药对象将来自于文献。为了有效地发现文本之间存在的大量潜在知识，本文提出一种基于主谓宾(S, P,O)结构的最小知识单元表示及其语义推理的中文文本知识发现方法。在文本信息检索领域中也开始了研究文本句子结构的主谓宾信息，如文献[11]用提问和文献中的 SVO 结构改进全文搜索方法，把 SVO 结构切分成几个关键词，用关键词出现的顺序表达语义关系。文献[12]分析文本中句子的主谓宾结构(SVO)，利用两个句子间主谓宾成分之间的语义相似性提高检索精度。本文的方法与已有方法的区别在于，把中文句子的主谓宾结构抽象为 6 种(SPO)模式，首先对文本进行词性标记，对句子进行因果特征识别和过滤，从句子中提取出具有实意动词的主谓宾最小知识单元，把具有(SPO)结构的知识单元作为语义互补推理单元，建立知识单元的数据库，然后通过知识单元之间的谓语动词的作用与反作用语义关系进行推理发现文献之间的潜在新知识。实验结果表明，本方法与目前已知道的各种文献知识发现方法相比，不仅克服了传统的基于关键词发现非相关文献知识方法带来的大量噪声问题，也解决了主谓宾结构的知识单元的提取的理论与方法，同时创立了语义互补推理的文献间潜在知识发现的方法。这种新方法大大提高了文献之间隐含关联的知识发现效率，将有助于跨学科研究的科学技术研究者发现交叉学科之间存在的隐含的新知识。

2 传统方法与我们的方法

2.1 Swanson 的文献知识发现方法

2001 年，Weeber[4] 等人将 Swanson 非相关文献发现的方法总结归纳为“两步法”模式，即形成假设和检验假设。第一步为疾病寻找新的治疗方法，或为药物寻找新的靶标，称为开放式发现方法(open)，发现过程可表示为 $C \rightarrow B \rightarrow A$。第二步以 A 和 C 为出发点，研究人员需要努力寻找共同的中介词 B。A 和 C 的联系越多，所做的假设越有价值。检验假设的过程称为闭合式的发现方法(closed)，闭合式的文献发现过程可表示为 $A \rightarrow B \leftarrow C$。开放式和闭合式文献发现方法可表示为图 1 所示。

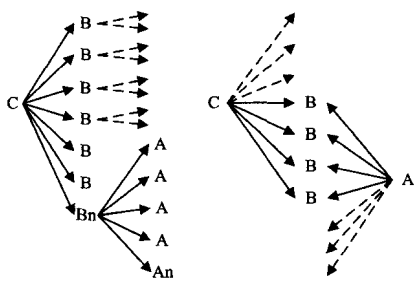

图 1 开放式和闭合式非相关文献发现方法

同时，Weeber 等人认为此前 Swanson 等人的发现方法均是基于单词、双字词和三字词的常规文献检索和验证，这不适应于目前越来越普遍的语义分析环境，众多语义仅仅用单词、双字词和三字词是无法概括和准确表达的。因此，Weebe 也提出了适合语义分析环境的基于多字词的发现方法，即所谓的基于“概念”的非相关文献发现。

2.2 知识单元语义推理的文献发现模型

我们方法的原理是首先抽取出文摘中结论部分的有效因果关系的 SPO 结构，建立带有语义关系的 SPO 知识单元库；2)通过 SPO 知识单元库，基于语义互补关系推理发现文献之间的隐含知识；3)我们的方法克服了以往方法由于语义丢失带来无关匹配的大量噪声问题，能减少人工参与的工作量。我们基于 SPO 结构的知识单元语义互补推理知识发现原理如图 2 所示。

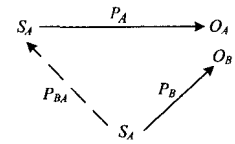

图 2 SPO 结构的知识单元语义互补推理知识发现原理

图 2 中，大写字母 S,P,O 分别为主语、谓语、宾语，A,B 分别为两篇不同文献的知识单元，分别表示为一个 SPO 结构三元组 $A(S_A, P_A, O_A)$，$B(S_B, P_B, O_B)$；当 $O_A = O_B$，$P_B = \overline{P_A}$，即 A,B 两个知识单元的宾语值相等，语义关系为互补时，则可能发现一条由语义推理 $S_B \rightarrow S_A$ 得到的一条不在同一文献中的新知识，$BA(S_B, P_{BA}, S_A)$。

3 SPO 知识单元挖掘

3.1 文献的 SPO 知识单元模型

2002 年，SHI Yu. Zhi 博士认为：凡不论哪一种语言，其句子的基本成分都是 S(主语)、V(谓语动词)和 O(宾语)。这 3 个基本成分的可能排列顺序有 6 种：SVO、SOV、OSV、VSO、VOS、OVS。采用 SVO 和 SOV 的语言最多，大约占整个人类语言的 90%。其它的就很少了，后两种几乎不见。汉语从古到今都是 SVO 语言，这在很大程度上决定了汉语语法系统的整体面貌，古英语是 SOV 型的，从 15 世纪之后变成了 SVO 语言。同样斯坦福大学 Kiparsky 教授根据对大量语言发展史的观察指出，世界上很多原来为 SOV 的语言，都变成或者正在变成 SVO，并指出利用 SVO 这种语序区别句子基本成分的语言，是人类语言最优化的结构，最符合人类的认知顺序。

基于词是概念的最小单位，句子是知识交流的最小单位的理论，本文研究以句子为单位的文本知识单元挖掘模型。借助于文献[13]的观点，首先将属性文法分为句法和语义两部分。句法部分由一个有限状态文法或上下文无关文法组成，语义部分由语义规则组成。

(1)形式化句子定义

定义 1 定义一个带属性的上下文无关串文法为四元组：

$$G = (V_N, V_T, P, S) \tag{1}$$

其中，V_N 是一个非终止符集，V_T 是一个终止符集，S 是起始符集，对于每一个 $X \in (V_N \cup V_T)$，存在着属性 $A(X)$ 的有限集，且每个 $A(X)$ 中的 D_a 具有一个 D_a 可能值的有限集或无限集；P 是产生式集，集中的产生式由语法规则和语义规则两部分组成。

定义 2 定义带属性的语法为联接关系组：

$$X_0 \rightarrow X_1 X_2 \cdots X_m \quad (2)$$

其中，$X_0 \in V_N$，且每一个 $X_i \in V_N$，$1 \leqslant i \leqslant m$，$a_i$ 与 a_j 两个符号间的联接关系为 $CAT(a_i, a_j) \in R(a_i, a_j)$。

定义 3 定义语义规则为语义函数组：

$$\begin{aligned} a_1 &\rightarrow f_1(a_{11}, a_{12}, \cdots, a_{1n}) \\ a_2 &\rightarrow f_1(a_{21}, a_{22}, \cdots, a_{2n}) \\ &\cdots \\ a_m &\rightarrow f_1(a_{m1}, a_{m2}, \cdots, a_{mn}) \end{aligned} \quad (3)$$

其中，$a_1, a_2, \cdots, a_n = A(X_0) \cup A(X_1) \cup \cdots \cup A(X_n)$，每一个 a_{ij} $(1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n)$ 代表 X_k $(0 \leqslant k \leqslant m)$ 的一种属性，f_i $(1 \leqslant j \leqslant n)$ 被称为语义函数。

(2)抽象化 SPO 知识单元定义

考虑到中文句子的谓语组成的复杂性，如：(癫痫/n)(持续/v 大/d 发作/v 导致/v)(急性/b 肺水肿/n —/m /q 例报告/n)，我们将汉语句子抽象为 6 种 SPO 句型，以此建立 SPO 结构的文本知识单元结构。

定义 4 定义文本知识单元(TKE)为一个三元组：

$$TKE = (S, P, O) \quad (4)$$

其中，S 为主语，P 为谓语，O 为宾语。P=(verb-adjective, adjective-verb, verb and subject predicate, Compound predicate, Predicate, verb)。由此基于 SPO 结构的文本知识单元可以表示为下面 6 种形式：

$TKE_1 \rightarrow f_1$(Subject, verb-adjective, object)

For example:(小张，穿上军装-很，精神)

$TKE_2 \rightarrow f_2$(Subject, adjective-verb, object)

For example:(孩子，不小心-跌了，一跤)

$TKE_3 \rightarrow f_3$(Subject, verb and subject predicate, object)

For example:(小虎，看了(这部)影片热泪，盈眶)

$TKE_4 \rightarrow f_4$(Subject, Compound predicate, object)

For example:(运动，使您身体，健康)

$TKE_5 \rightarrow f_5$(Subject, object, Predicate)

For example:(山姆，把橘子，给吃了)

$TKE_6 \rightarrow f_6$(Subject, verb, object)

For example:(小孩，吃，苹果)

3.2 SPO 知识单元挖掘算法

本文的算法采用以下步骤进行：首先，对文本进行预处理。采用中科院 ICTCALS 分词工具对文献中的句子进行分词，并加词性标记。接着，判断所选择的句子是否符合 6 种 SPO 结构。再次，识别出描述因果事件的 SPO 句。

对状态事件和因果事件的判断规则如下：

(1)因为主语是有意志的行为者或者是无意志的事物，如果谓语动词不能使宾语发生状态变化，那么 SPO 就只能表达一个知识单位的状态事件。

(2)如果谓语动词具有实义，即 O 成为结果宾语，主语成为该结果宾语的原因，这时的 SPO 结构就描述一个知识单位的因果事件。

在这里，我们先建立一个动词字典进行处理和人工检验，在此基础上加入机器学习方法做一个补充。通过以上的算法步骤就可以把文献中的句子分解成所需要的 SPO 结构知识单元。下面给出生成 SPO 结构知识单元的伪代码。

```
For 读有词性标记语句的第一条语句到最后一条语句
  对语句进行分词标记(名词用/n 表示，动词用/v 表示)"/"。
  If 语句中有动词和名词
    寻找该语句中第一个动词和最后一个动词的位置
    If 第一个动词前有名词
      第一个动词前所有字符串作为主语；
      将其名词作为分析出的主语，谓语则为第一个动词和最后一个动词的字符串，宾语为最后一个动词后的字符串。
    Else 在第一个动词后的所有字符中找到名词
        把名词作为主语，谓语为该名词和最后一个动词的字符串，宾语为最后一个动词后面的字符串。
    End
  Else 删除该句子
  End
End for
```

3.3 生成 SPO 结构的知识单元库

通过上述的算法把中文句子分解成主语、谓语、宾语(S, P, O)三元组结构的因果关系的知识单元，然后把这种结构自动地添加到 S, P, O 3 个字段的数据库中进行保存，从而实现了把非结构化的文本数据处理成结构化的数据库。由此实现基于主谓宾句子结构的 SPO 语义检索模型，改变了传统的向量空间信息检索语义丢失带来的大量噪声。SPO 克服了 SVO 单一动词语义不足和处理上的麻烦，同时为基于语义推理的知识发现建立了基础。以不相关的医学文献为例给出利用本文的 SPO 结构知识单元抽取算法得到的结果。SPO 结构知识单元库如表 1 所列。

表 1 SPO 知识单元库结构

NO	Subject	Predicate	Object
1	/w 压力/n	引起/v	偏头痛/n
2	/w 压力/n	导致/v	镁/n(/w 金属元素/n)/w 流失/n
3	//w 钙/n 通道/n 阻断剂/n	预防/v	某些/r 偏头痛/n
4	/w 镁/n	是/vl	天然/a 的/u 钙/n 通道/n 阻断剂/n
5	/w 偏头痛/n 中/n	涉及/v	撒/n 布/n 皮质/n 凹陷/nSCD/ws
6	/w 高/a 强度/n 镁/n	抑制/v	SCD/ws
7	/w 偏头痛/n 患者/n	有/v	高/a 血小板/n 聚集/n 功能/n
8	/w 镁/n	能阻/v	血小板/n 聚集/n 功能/n
9	/w 长期/nt 镁/n 缺乏/n	有/v 致痛/v	作用/n
10	/w 胰岛素/n	可/vu 促进/v	镁/n 的/u 吸收/v

4 SPO 语义互补推理的知识发现

4.1 SPO 知识单元的知识发现模型

SPO 是一个含有语义关系的知识单元，借助于 SPO 知识库的存储和检索，我们就可以通过语义关系的推理操作来发现隐含的新知识。

设 S 是一个 SPO 知识单元的主语，O 是 SPO 知识单元的宾语，P 是 SPO 知识单元的谓语。

定义 5 语义互补推理是一种根据多个 SPO 知识单元中实义动词间的互补关系的递推，即寻找两个知识单元 *SPO*-1 与 *SPO*-2 之间存在的"作用"与"反作用"的互补语义关系，递推出一个新的知识单元 *SPO*-3 的过程。

例如：已知 S_1，得到 O_1，且 S_1 与 O_1 存在"作用 P_1"语义关系，寻找一种与 O_1 存在"反作用 P_2"语义关系的 S_2，如果

找到了 P_2，则完成了由 S_2 到达 S_1 的“推理 P_3”语义关系。

4.2 SPO 知识单元的隐含知识发现算法

输入查询名词，记为 X
For i=1：13
  在字段 3 中查询 X
  得到包含 X 的行，并抽取相应的字段 1 中的元素记为 Y；
  判断字段 2 中 Y 与 X 的语义关系；
  If X 由 Y 引起
    得到因果关系的 SPO 三元组知识单元
  Else 得到 Y 作用于 X 的反作用关系的 PO 三元组知识单元
End
For i=1：13
  在字段 1 中查询 X；
  得到包含 X 的行，并抽取相应的字段 3 中的元素记为 Z；
End
For i=1：13
  在字段 3 中查询 Z；
  得到包含 Z 的行，并抽取相应的字段 1 中的元素记为 D；
End
For i=1：13
  在字段 1 中查询 D；
  得到包含 D 的行，并抽取相应的字段 3 中的元素记为 E；
End
FOR I=1：E
  在字段 2 中查询反作用语义关系
  If 存在反作用关系
    得到 E 与 D 构成的三元组知识元
  End
End
得到 Z 到 X 的知识推理

4.3 SPO 语义互补推理的知识发现系统实现

(1)SPO 知识单元数据库推理发现系统

利用本文第 3.2 节提出的 SPO 知识单元挖掘算法，我们在 Microsoft SQL Server 2008 数据库系统上建立了知识单元数据库；利用本文 4.2 节提出的 SPO 知识单元的隐含知识发现算法，我们建立了基于 SPO 知识单元数据库推理的隐含知识发现系统，如图 3 所示。图 3 系统展示了一个与“高血压升高的相关原因，通过推理直接得到降低高血压的几种方法(药物或管理)的结果”。

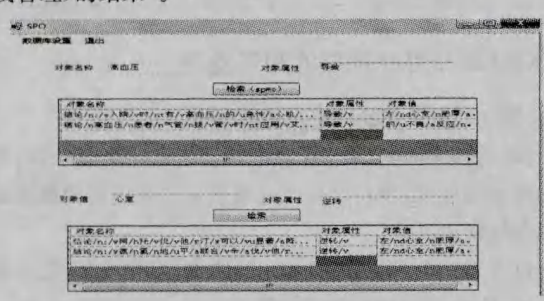

图 3 SPO 知识单元推理的隐含知识发现系统

(2)系统的语义推理发现分为 3 个部分实现

第一部分检索 $S->P->O$ 知识单元：

首先检索“高血压”“导致”的其他疾病的 SPO 知识单元。从 S 字段(对象名称)输入“高血压”，且在 P 字段(对象属性)输入“导致”，获得 2 条“高血压”“导致”其他疾病的 SPO 知识单元。即 $S->P->O$ 知识单元(23772，23882)。结果如表 2 所列。

表 2 检索 $S->P->O$ 知识单元

S1	P1	O1	文摘号
高血压/n 急性/a 心肌/n 梗死/v 患者/n	导致/v	左/nd 心室/n 肥厚/a	23772
高血压/n 患者/n 气管/n 拔/v 管/v 应用/v 艾/n 司/n 洛/x 尔/r 能/vu 明显/a 降低/v 应/vu 激/v	导致/v	的/u 不良/a 反应/n	23882

第二部分检索 $O->P->S$ 知识单元：

检索“逆转”“左心室肥厚”的 SPO 知识单元。从 O 字段(对象值)输入“左心室肥厚”，且在 P 字段(对象属性)输入“逆转”，获得 3 条能“逆转”“左心室肥厚”的 SPO 知识单元。即 $O->P->S$(23531，31034)。结果如表 3 所列。

表 3 检索 $O->P->S$ 知识单元

S2	P2	O2	文摘号
阿/h 托/v 伐/v 他/r 汀/x 可以/vu 显著/a 降低/v 血压/n 和/c	逆转/v	左/nd 心室/n 肥厚/a	23531
氨/n 氯/n 地/u 平/a 联合/v 辛/a 伐/v 他/r 汀/x 能/vu 有效/a 降压/v 和/c	逆转/v	左/nd 心室/n 肥厚/a	31034

第三部分 $S2->S1$ 语义推理的新知识发现：

由于 $O1=O2$，且 $P2$ 与 $P3$ 语义互补，可由推理 $S2->S1$，获得 $S2$ 抑制，降低 $S1$。结果如表 4 所列。

表 4 $S2->S1$ 语义推理的新知识发现

S2	P3	S1	推出
采用/v 阿/h 托/v 伐/v 他/r 汀/x 联合/v 厄/x 贝/n 沙/n 坦/	降低/v	高血压/n 的/u 急性/a 心肌/n 梗死/v	

5 实验分析

5.1 实验结果比较分析

(1)文献[14]的传统方法

文献[14]实验数据采用中国知网、维普期刊和万方文献数据库等网上公开的中医药临床文献数据库，从中选取 500 篇文摘实验。文献[14]的做法是将文摘中的目的部分分解为目的(病名，药物)，将结论部分分解为结论(病名，药物)，如表 5 所列。

表 5 文献[14]的抽取方法

目的-病名	目的-药物	结论-病名	结论-药物
慢性阻塞性肺疾病	补肺活血化痰法	COPD 急性加重	补肺活血化痰中药
常年性变应性鼻炎	鼻炎固本方	常年性变应性鼻炎	鼻炎固本方

文献[14]作者认为自己的方法实验结果精确度过低，在文献数量为 100 篇和 200 篇时，发现的关联数为 0，即使在 500 篇时，发现的有意义的关联也只有 2 篇，也就是准确率只有 0.004，所以准确率较低，且需要大量人工检查。

下面我们通过文献[14]给出的抽出表 5 的 2 篇文摘的结论句进行分析。1)结论：补肺活血化痰中药可有效减少 COPD 急性加重次数，改善临床症状。2)结论：论鼻炎固本方治疗脾肾阳虚型常年性变应性鼻炎有较好的疗效。

可以看出：改造后的传统方法有两个不足：1)噪声很大，即 500 篇结论句中只抽出 2 条关联句；2)抽出的关联词还只

是文摘结论中的病名、药名；3)该方法难以发现分散在不同一篇文献之间的隐含关系，所需人工参与量非常大，实现起来很困难。

(2)本文方法显性知识发现效率分析

1)标题和文摘 SPO 知识单元抽取效率果比较

我们对 1000 条"高血压"标题的实验结果发现，文献标题不适合于基于 SPO 的知识单元抽取方法。实验表明从标题抽出的 SPO 结构是一种报道性事件模型，而不提供事实性因果性结论。相比之下，文摘的结论部分给出了事实性的结论，是有效 SPO 知识单元抽取较好的来源之一。例如：对同一篇期刊文献的标题和文摘抽取结果的对比如表 6 所列。

表 6　标题与文摘的 SPO 知识单元表示比较

Title-1			
S	P	O	ID
氨/n 氯/n 地/u 平/a 联合/v 阿/h 托/v 伐/v 他/r 汀/x 治疗/v 老年/nt 高血压/n 合并/v 高血脂/n 的/u	临床/v	疗效/n	27456
Abstract-2			
S	P	O	ID
氨/n 氯/n 地/u 平/a 联合/v 阿/h 托/v 伐/v 他/r 汀/x 钙/n 治疗/v 高血压/n 伴/v 高血脂/n	具有/v	疗效/n 佳/a	26105

2)标题和文摘 SPO 知识单元抽取数量比较

文献标题和文摘 SPO 知识单元的抽取效率比较如表 7 所列。

表 7　标题和文摘 SPO 知识单元的抽取效率

实验对象	抽取文献数	抽出 SPO	抽取效率	因果 SPO	有效率
标题 1（高血压）	1000	262	26.2%	1	0.4%
文摘 1（高血压）	679	362	53.3%	330	48.6%
文摘 2（中医药）	450	250	55.5%	213	47.3%
文摘 3（偏头痛）	368	163	44.3%	135	36.7%

(3)本文方法对文献间隐含知识发现效率分析

文献间的知识发现是指利用基于有因果结论关系 SPO 知识单元的语义推理的结果。因此文献间的发现效率与多个因素有关：1)与文献本身的研究和表达水平密切相关；2)与 SPO0 知识单元的抽取的准确率有关；3)与 SPO 知识单元语义推理算法有关；4)与发现新知识的判断者的知识水平有关。第 2 种因素前面已经分析过了，这里主要分析第 3 种推理算法因素。推理算法效率主要从方便性和准确性考察。

方便性好：将文献之间的比较转换为 SPO 知识单元间的语义关系比较，大大提高了比较的速度和方便性。假设一个人一天能快速阅读 50 篇科技文献，要在 1000 篇文献之间通过比较发现文献间新的互补性知识将需要 20 天时间。而使用我们的数据库结构的 SPO 知识元推理，可能只需要 1～2 天时间就够了，而且阅读压力大大减小。

准确性高：由于通过数据库结构的 SPO 3 个字段查找，使得要发现的对象范围大大缩小，比如先通过疾病字段 O，找到与该疾病有关的所有药物 S，再通过语义字段 P，了解药物与该疾病的使动因果关系 $P1$。在此信息关系的基础上，通过药物字段 S 的扩大检索，发现了药物还有的其他功能和语义关系 $P2$，通过 $P1$ 与 $P2$ 的互补语义关系，即可发现新知识的推测。

结束语　本文的贡献在于①提出将文献分解成 SPO 结构的最小知识单元表示，克服了传统的以关键词表示文献知识碎片带来的语义丢失所产生的大量噪声和虚假组合的干扰问题。②把中文句子的主谓宾结构抽象为 6 种 SPO 结构模式，对句子进行因果特征识别和过滤，解决了中文句子多个动词的主谓宾构成的最小知识单元的自动获取问题。③将具有主谓宾结构(SPO 结构)的知识单元作为语义互补推理单元。④将 SPO 存储成 3 个字段组成的语义数据库，实现了从 S 和 O 两个字段检索，并利用 P 的语义关系推理发现文献间隐含新知识的方法。实验证明本文提出的带有语义关系的 SPO 知识单元比传统的关键词发现文献知识效率高，且文摘比标题对发现文本知识更有价值，本文提出的方法与已有方法相比有方便、快速、效率高等优点。本方法的实验只是选择了 3 个单一专题文献，还不是多学科文献，就已看出本文的方法鼓舞人心的价值。接下来的工作将进一步完善抽取算法的效率和多学科文献的联合实验，进一步提高跨学科文献间的隐含知识的发现效率。

参 考 文 献

[1] Swanson, Oil D F. Raynaud's syndrome, and undiscovered public knowledge[J]. Perspectives in Biology and Medicine, 1986 (30): 7-18

[2] http://www.asis.org/Bulletin/Mar-01/swanson

[3] Chen Z. Let Documents Talk to Each Other: A Computer Mod for Connection of Short Documents[J]. Journal of Documentation, 1993, 49(1): 44-54

[4] Weeber M, Kkin H, de Jong-van den Berg. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries[J]. AmSoc Inf Sci Technol, 2001, 52(7): 548-557

[5] Illhoi Y. Semantic text mining and its application in biomedical domain[D]. Drexel Univeristy, 2006

[6] Tsuruoka Y, Tsujii J, Ananiadou S. Facta: a text search engine for finding associated biomedical concepts Bioinformatics[J]. Bioinformatics, 2008, 24(21): 2559-2560

[7] Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery[J]. Journal of Biomedical Informatics, 2009, 42(4): 633-643

[8] Cohn T, Schvaneveldt R, Widdows D. Reflective random indexing and indirect inference[J]. Journal of Biomedical Informatics, 2010, 43(2): 240-256

[9] 周峰，林鸿飞，杨志豪. 基于语义资源的生物医学文献知识发现[J]. 情报学报，2012，31(3)：268-274

[10] Roger H. Text Mining: Getting more Value from Literature Resources[J]. Drug discovery today, 2005, 10(06): 377-379

[11] Gao Ming, Wang Ji-cheng. Semantic Search Based on SVO Constructions In Chinese[C]//2009 International Conference on Web Information Systems and Mining. IEEE, 2009: 213-216

[12] 黄承慧，印鉴，侯防. 一种基于主谓宾结构的文本检索算法[J]. 计算机科学，2010，37(9)：173-176

[13] 戴汝为. 语义句法模式识别方法及其应用[M]. 西安：西安交通大学出版社，2011.7

[14] 方纯洁，王波，罗杰，等. 基于信息抽取的中医药文献知识发现[J]. 浙江中医药大学学报，2012，36(1)：88-90