基于形式概念分析的信息检索算法

赵玲

（大连海事大学信息科学技术学院 大连 116026 ）

摘 要 随着网上的信息不断的增长，信息检索算法变得至关重要。实际上，目前用于信息系统设计的方法仍然不能满足用户的需求，无论是从性能（查准率和查全率）还是响应时间。在本文中，提出一种基于形式概念分析的检索算法。事实上，信息检索是形式概念分析的直接应用。在此背景下，利用 FCA 提供的理论基础设计一种新的信息检索方法。

关键词 形式概念分析；信息检索算法；连续查询

中图法分类号 TP 301

Formal Concept Analysis regarding information retrieval algorithm

Zhao Ling

(Department of Information Science and Technology, Dalian Maritime University, Dalian, 116026*, China)*

Abstract With the exponential increase in the quantity of information circulating on the Internet, an evolution of information-retrieval systems becomes paramount. Indeed, current approaches for information systems design remain unable to meet the needs of users, either in performance (precision and recall) or response time.In this paper, we propose a information-retrieval algorithm based on formal concept analysis. In this context, we exploited the theoretical basis provided by the FCA to design an approach for information retrieval.

Key words formal concept analysis;information retrieval algorithm;conjunctive query

1 引言

形式概念分析(Formal Concept Analysis，FCA）是由 R.Wille 于 1982 年提出的一种从形式背景进行数据分析和规则提取的强有力工具，形式概念分析建立在数学基础之上，对组成本体的概念、属性以及关系等用形式化的语境表述出来，然后根据语境，构造出概念格(concept lat-tice),即本体，从而清楚地表达出本体的结构。这种本体构建的过程是半自动化的，在概念的形成阶段，需要领域专家的参与，识别出领域内的对象、属性，构建其间的关系，在概念生成之后，可以构造语境，然后利用概念格的生成算法 CLCA，自动产生本体。形式概念分析强调以人的认知为中心，提供了一种与传统的、统计的数据分析和知识表示完全不同的方法，成为了人工智能学科的重要研究对象，在机器学习、数据挖掘、信息检索等领域得到了广泛的应用。

文档索引是信息检索过程中非常重要的任务。索引给定文档集合以提取有意义的词汇标记开始。这些术语可以以详尽且明确的方式表示文档的内容。在以往的研究中，已经开发了许多模型用于从文档数据库中提取相关知识。

在提取相关术语之后，对它们的文档和术语之间的关系进行建模。在索引期间，需要满足几个约

收稿日期：2016-11-15；最终修改稿收到日期：2016-11-15 . 本课题得到国家自然科学基金(70471003)资助.赵玲，女，1993年生，硕士，主要研究方向为智能信息处理，E-mail: 2495662643@qq.com.

束。良好的索引系统必须在短时间中确保有效的信息检索。在实践中，信息检索系统设计者面临着随时间演变的问题。这个问题与大量处理的数据有关。

在以往的研究中，已经提出了几种方法来开发信息检索系统。这些方法可以总结为三类：

布尔模型：这个模型是最古老的；它基于文档和查询的逻辑表示。

矢量空间模型（VSM）:在这个模型中，文档和查询根据它们的术语在矢量空间中作为矢量呈现。

概率模型：这是基于概率理论的数学模型。

一般来说，基于 FCA 的信息检索模型可以分为三类：CLR（基于概念网格的排名）方法，细化和组织方法以及基于分类的推理方法。

2 形式概念与形式背景

2.1 形式背景

形式概念分析的首要工作便是建立形式背景。

定义 1 形式背景:该形式背景是一个三元组 $F=(D,T,I)$ 其中:

D:对象的集合；

I:属性的集合；

I:D 和 T 之间的二元关系,并且 $I\in D\times T$；

其中,*I* 是 F 的关联关系,让 $d\in D$ 和 $t\in T$,则二元关系 $(d,t)\in I$ 表示 *d* 包含 *t*,还可以说成是 *d* 满足 t。

定义 2 求导算子，也被成为充分性算子。求导算子被认为是形式分析的基础。给定 $X\in D$ 和 $Y\in T$，用 X^{Δ}（等式 1）表示对象集合满足 X，用 Y^{Δ}（等式 2）表示属性集合满足集合 Y。

$$X^{\Delta}=\{y\in T\mid X\subseteq I(y)\}\quad(1)$$

$$Y^{\Delta}=\{x\in D\mid Y\subseteq I(x)\}\quad(2)$$

直观的说，X^{Δ} 是 X 的所有对象共有的属性集，Y^{Δ} 是共享 *Y* 的所有属性的对象的集合。

2.2 形式概念

定义 3 形式概念：形式概念是一对 $<X,Y>$，使得 $X^{\Delta}=Y$ 和 $Y^{\Delta}=X$。是一种顺序关系，表示为≤，定义在所有概念的集合上。 有以下等式（等式 3）：

$$<X_1,Y_1>\leq<X_2,Y_2>\Leftrightarrow\Leftrightarrow X_1\subseteq X_2,Y_2\subseteq Y_1\quad(3)$$

3　概念格

3.1 概念格的基本概念

概念格是 FCA 的核心数据结构。概念格的每个节点是一个概念，由外延和内涵组成。外延是概念所覆盖的实例;而内涵是概念的描述，是该概念所覆盖实例的共同特征。概念格可以通过其 Hasse 图生动简洁地体现概念之间的泛化和例化关系。概念格结构模型是形式概念分析理论中的核心数据结构。其本质上描述了对象和特征之间的联系，表明了概念之间的泛化与例化关系。这种概念格构建的过程是半自动化的。

定义 4 设 $<k,\leq>$ 为偏序集，$D\in K$，*a* 为 *K* 的任一上界，若对 *D* 的所有上界 y 均有 $a\leq y$，则称 a 为 *D* 的最小上界，即上确界。同样，若 *d* 为 *D* 的任一下界，若对 D 的所有下界 *z* 均有 $z\leq d$，则称 *d* 为 *D* 的最大下界,即下确界。

定义 5 设 $<k,\leq>$ 为偏序集，如果 K 中任意两个元素都有最小上界和最大下界，则称 $<k,\leq>$ 为格。

定义 6 对于形式背景 $F=(D,T,I)$ 存在唯一的一个偏序集 $<k,\leq>$ 与之对应，并且该偏序集的子集的上确界与下确界都存在，这个偏序集产生的格结构称为概念格。

4 信息检索算法中的连续部分

4.1 连续查询模型

定义 7 令 $E=(\{d\},td)$ 是形式上下文 $F=(D,T,I)$ 中的条目。E 是由两组组成的对。第一

个集合（$\{d\}$）包含单例（文档 *d*），第二个（t_d）包含对应于文档 d 的术语。给定一个对 $E_Q=(d_Q, t_Q)$，其中 d_Q 是满足连接查询 Q 的文档的集合，t_Q 是 Q 的项集。当且仅当 $t_Q \subseteq t_d$，即文档 *d* 与查询 Q 相关，t_d 包含 t_Q 的所有项。

令 Q 是由 n 项的连接组成的查询　（公式 4）

$$Q = t_1 \Lambda t_2 \Lambda ... \Lambda t_{n-1} \Lambda t_n \qquad (4)$$

令 $D_Q \in D$ 是满足 Q 的文档集。令 $C =< X, Y >$ 是属于 $F(D,T,I)$ 的最一般的概念，使得 $\{t_1, t_2, ..., t_{n-1}, t_n\} \subseteq Y$，则 $D_Q = X$。提出以上定义（7），其定义是文档与联合查询的相关性。

4.2 算法提出

该算法基于两个主要想法：首先是从网格中的最具体的概念（底部概念）开始，在网格中识别最一般的概念，使得其包含所有查询项。第二是为了简化搜索空间，忽略其不包含查询项概念的祖先。在以下命题中制定方法论。

定理 1：如果一个概念不满足连接查询，那么有必要检查它的祖先。

证明：如果 $C2 =< X2, Y2 >$ 是 $C1 =< X1, Y1 >$ 的 $C1 \le C2$ 的祖先，这意味着 $Y2 \subseteq Y1$。 如果 $\{t_1, t_2, .., t_{n-1}, t_n\} \not\subset Y_1$ 然后 $\{t_1, t_2, .., t_{n-1}, t_n\} \not\subset Y_2$，这表明不必检查 *C2*。

该算法（算法 1）通过检查最具体概念（底部）的意图是否包含查询项来开始搜索。如果底部概念的意图不包含 Q 的所有项，则结束治疗而不发现满足查询的任何文档（在这种情况下：$Rd = \{\phi\}$）。如果不是，继续探索它的直接父亲。 如在上一步骤中，如果发现没有包含查询项的意图，则结果是底部概念的范围（$R_d = \{X_\beta\}$）。 重复相同的处理，直到检查格子的所有概念，通过跳过那些不满足查询的祖先来减少每次搜索空间。

Data: $F = (D, T, I)$, $B(D, T, I)$, $Q = t_1 \wedge t_2 \wedge ... \wedge t_{n-1} \wedge t_n$
$T_Q = \{t_1, t_2, ..., t_{n-1}, t_n\}$
ψ: set of concept to explore
β: Lattice Bottom // most specific concept in the lattice
Result: R_d: set of documents satisfying the query Q
begin
1　$R_d \longleftarrow \{\emptyset\}$
2　$\psi \longleftarrow \{\beta\}$
　　for $C \in \psi$ **do**
3　　**if** $T_Q \subseteq intent(C)$ **then**
4　　$Clean(R_d)$
5　　$R_d \longleftarrow extent(C)$
6　　$\psi \longleftarrow \psi \cup Fathers(C)$
7　　$\psi \longleftarrow \psi \setminus \{C\}$
　　else
8　　$\psi \longleftarrow \psi \setminus \{C\}$
9　　delete $Ancestors(C)$ from the lattice
　　end
　end
end

4.3 算法评估

信息检索算法中的的结合部分是良好的，遵守定义 7 给出的相关性标准。

证明：

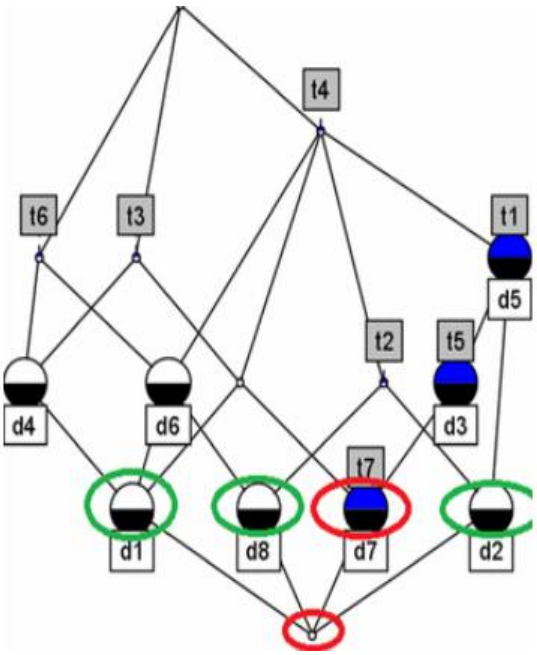
令 $E = (\{d\}, td)$ 是形式上下文 F 中的数目。d 与涉及 $tQ \subseteq td$ 的 Q 相关。有：$d \in D$ 意味着 $\exists C =< extentC, intentC > \in B(D,T,I)$，使得 $td \subseteq intentC$ 和 $d \in extentC$。 信息检索算法的连接组件探索 $F(D,T,I)$（包括 C）中的所有概念的意图，证明由算法检索到 *d*。定理 1 和 2 允许陈述以下定理。

定理 3 信息检索算法的连接分量是健全和完整的，遵守由定义 7 给出的相关性标准。

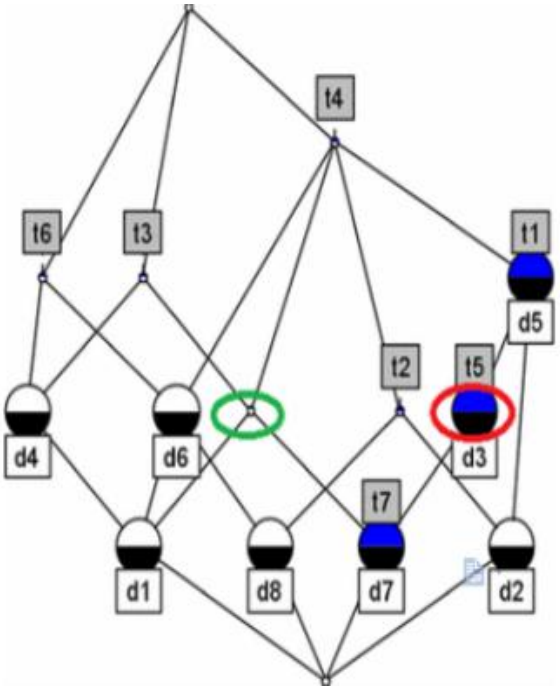
4.4 举例

考虑以下查询：$Q = t_1 \Lambda t_4 \Lambda t_5$.

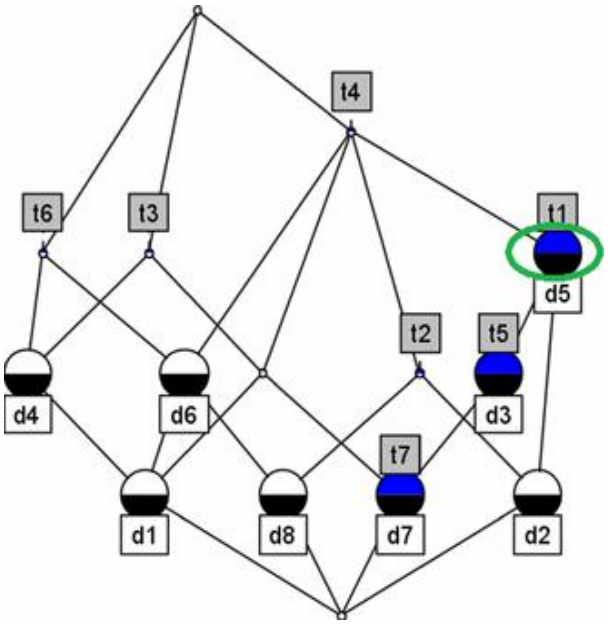
算法 1 通过探索最具体的概念格（β）开始。显然，β 的意图必然包含所有查询项。在该步骤中，$R_d = \{\phi\}$。概念 β 具有四个父亲（参见图 2）。找到一个单一的父结点，其意图包含查询词（在图 5 中以红色圈出）。在这种情况下，$R_d = \{d_7\}$。



此后，该算法研究找到底层概念的父亲。这个概念有两个父亲。但是唯一的概念在其意图中包含查询项。在该步骤中，$R_d = \{d_3, d_7\}$。



下一步是查找发现概念的父亲，这个概念的意图不包含 t_5。然后，算法在此阶段停止。因此，最终得到 $R_d = \{d_3, d_7\}$。



5 结束语

在本文中，提出了一种基于形式概念分析的信息检索的新算法。为了降低算的复杂性，从概念格中的概念之间的包含关系中分析。

理论证明和实验研究表明，根据执行时间测试该模型的性能非常好。

作为未来的工作，应该尝试扩展该方法覆盖其他类型的布尔运算符，如否定运算符。实际上，该方法仅适用于连续查询。可以处理包含不同运算符的查询的通用算法在通过概念网格的信息检索的领域中可能非常重要。作为第二潜在扩展，尝试使用语义资源（同义词或本体）建立网格的概念之间的语义关系。

参 考 文 献

[1]. Baranyi P, Gedeon TD, Koczy LT　Intelligent information retrieval using fuzzy approach. In: Systems, man, and cybernetics, 1998. 1998 IEEE international conference on, vol 2, 2010

[2]. Bordogna G, Pasi G Flexible querying of structured documents. In: Larsen H, Andreasen T, Christiansen H, Kacprzyk J, Zadrony S (eds) Flexible query answering systems, volume 7 of advances in soft computing. Physica-Verlag HD, 2012

[3]. Boughanem M, Loiseau Y, Prade H (1992) Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In: Proceedings of the third international conference on adaptive multimedia retrieval: user, context, and feedback, AMR'05. Berlin, Heidelberg, 2006. Springer, 2012

[4]. Callan J, Croft WB, Harding SM The inquery retrieval system. In: Proceedings of the third international conference on database and expert systems applications. Springer,2010

[5].Jon D Dvdsleuth: a case study in applied formal concept analysis for navigating web catalogs. In: Priss U, Polovina S, Hill R (eds) Conceptual structures: knowledge architectures for smart applications, vol 4604., lecture notes in computer science. Springer, Berlin, 2007

[6]Wille R (1982) Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival I (ed) Ordered sets, vol 83, NATO advanced study institutes series, Springer, Dordrecht, pp 445–470

Zhao ling born in 1993

E-mail:2495662643@qq.com, The main research direction is intelligent information processing

Background

During the last three decades, formal concept analysis (FCA) became awell-known formalism in data analysis and knowledge discovery because of its usefulness in important domains of knowledge discovery in databases (KDD) such as ontology engineering, association rule mining, machine learning, as well as relation to other established theories for representing knowledge processing, like description logics, conceptual graphs, and rough sets. In early days, FCA was sometimes misconceived as a static crisp hardly scalable formalism for binary data tables. In this paper, we will try to show that FCA actually provides support for processing