An Application of Text Mining: Bibliographic Navigator Powered by Extended Association Rules

Minoru Kawahara and Hiroyuki Kawano

Ying Yuan

Nov 27, 2000

Outlines

- Problem Definition
- Weighted Mining Association Algorithm
- Association Rules Using Several Attributes
- Performance

Problem Definition

- Subject
  - Bibliographic database
    - The INSPEC database that contains 2,085629 bibliographic titles
    - Data form – SGML (Standard Generalized Markup Language)
      - SGML declaration - describing rules of current document such as a character set or characters used as control characters.
      - DTD (Document Type Definition) - defines structure of the document such as tags and relationships between entities
      - A document instance - written in text

Problem Definition (Cont.)

- Subject(cont.)

Problem Definition (*cont.*)

- As DTD, we define items corresponding to all tags used in the INSPEC data

```
<Biblio>
<AccessionNumber>5512630</AccessionNumber>
<RecordType>02</RecordType>
<CopyrightStatement>Copyright 1997, IEE
</CopyrightStatement>
<Title>Data mining with composite events
 based sampling in a dynamic environment
</Title>
<Author>Kawano, H. Hasegawa, T.</Author>
<Abstract>Data mining, or knowledge
 discovery in databases, is the
                :
</Biblio>
```


Problem Definition (*Cont.*)

- Association Rule Mining:
  - Given two values *support* and *confidence,* and a set of transactions, an association rules is an expression $X \Rightarrow Y$

$$support(\mathcal{K}_c) = \frac{|\mathcal{T}_c|}{|\mathcal{U}|}.$$

$$confidence(\mathcal{K}_c) = \frac{|\mathcal{T}_c|}{|\mathcal{T}_g|}.$$

Weighted Mining Association Algorithm

- Definitions

$\mathcal{G}$	:	a set of keywords given in a query.
$\mathcal{O}$	:	a set of the keyword sets derived from $\mathcal{G}$ as association rules in the database.
$\mathcal{T}_g$	:	a set of tuples which contain $\mathcal{G}$ in the database.
$\mathcal{T}_a$	:	a set of tuples which contain any keywords in $\mathcal{G}$ in the database.
$\mathcal{K}_a$	:	a set of all keywords in $\mathcal{T}_a$.
$\mathcal{K}_c$	:	any combinations of $\mathcal{K}_a$.
$\mathcal{T}_c$	:	a set of tuples which contain both $\mathcal{G}$ and $\mathcal{K}_c$.
$\mathcal{U}$	:	the set of all tuples in the database.

Weighted Mining Association Algorithm (*Cont.*)

- Definition of support

$$support(\mathcal{K}_c) = \frac{W(\mathcal{K}_c)}{W_a},$$

$$W(\mathcal{K}_c) = \sum_{\mathcal{T}_c} \min_{\mathcal{K}_c} w_{ij},$$

$$W_a = \sum_{\mathcal{T}_a} \max_{\mathcal{K}_a} w_{ij}.$$

Weighted Mining Association Algorithm (*Cont.*)

- Definition of confidence

$$confidence(\mathcal{K}_c) = \frac{|\mathcal{T}_c|}{|\mathcal{T}_g|}.$$

Association Rules Using Several Attributes

- Categorizing attributes

Table 1. Categorization of attribute types.

Category	Meaning
Characteristic	This includes keywords which show the characteristics of a bibliography. These values are provided by the authors, the publishers and the database editors. ex. Title, Keyword
Description	This includes sentences, phrases and words which describe the contents of a bibliography. ex. Abstract, Table of contents, Index.
Supplement	This shows the supplement information of a bibliography. ex. Author, Publisher, Conference, ISBN

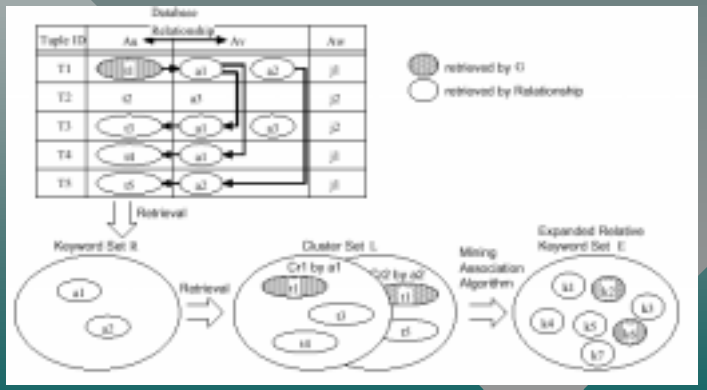
Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces
  - Statistics from INSPEC database

Category	Keyword space	Average
Characteristic	Small	Title 11 Keyword 30
Description	Small ~ Large	Abstract 183
Supplement	Small	Author 3

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Using relationship between attributes



Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Using relationship between attributes (cont.)

$$\mathcal{R} = \{a_1, a_2\}$$

$$t_1 \rightarrow \{a_1, a_2\} \rightarrow \{\{t_1, t_3, t_4\}, \{t_1, t_5\}\}.$$

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Algorithm

Algorithm 1

Input:

A keyword set $\mathcal{G}$ given in the input query,
A characteristic type of attribute A_c,
A characteristic type of attribute A_u, and a attribute A_v related to A_u.

Output:

An expanded keyword set $\mathcal{E}$

Association Rules Using Several Attributes (*Cont.*)

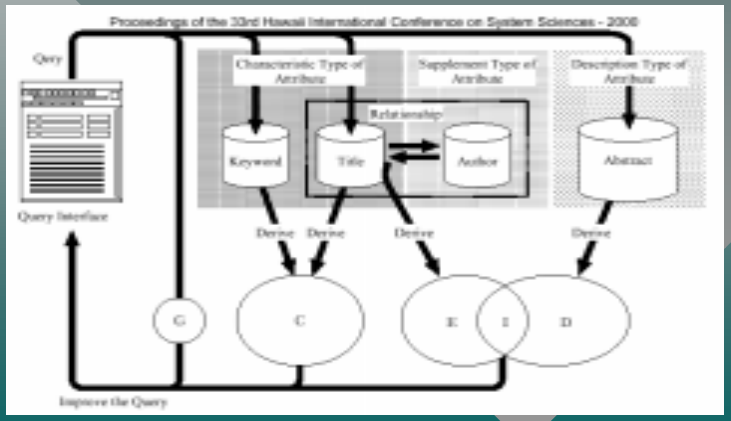
- Expanding keywords spaces (cont.)
  - Algorithm (*cont.*)

Method:

1. Select tuples which contain $\mathcal{G}$ in A_u and construct a tuple set T_g.
2. Gather the keywords from A_v in T_g and construct a keyword set $\mathcal{R}$.
3. Select tuples which contain a keyword in $\mathcal{R}$ and construct a tuple set T_r.
4. Gather the keywords from A_u in T_r and construct a cluster C_r.
5. Repeat Step 3 and Step 4 for each keyword in $\mathcal{R}$ and store the clusters into a cluster set $\mathcal{L}$.
6. Apply our mining association algorithm on $\mathcal{L}$ and derive a keyword set $\mathcal{E}$.

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Using relationship between *several* attributes



Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Algorithm

Algorithm 2

Input:
A keyword set $\mathcal{G}$ given in the input query,
A characteristic type of attribute A_c,
A characteristic type of attribute A_u, and a attribute A_v related to A_u,
A description type of attribute A_d.

Output:
An relative keyword set $\mathcal{O}$

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Algorithm (cont.)

Method:

1. Derive a keyword set $\mathcal{C}$ from A_c.
2. Derive a expanded keyword set $\mathcal{E}$ from the relationship between A_u and A_v applying the algorithm 1.
3. Derive a keyword set $\mathcal{D}$ from A_d.
4. If the intersection $\mathcal{I}$ of $\mathcal{E}$ and $\mathcal{D}$ produces an empty set, then the minimum *support* threshold $Minsup$ and the minimum *confidence* threshold $Minconf$ are lowered.
5. Go to step 1 if $Minsup$ and $Minconf$ are not lower than the limits given by the system administrator.
6. Calculate $\mathcal{O} = \mathcal{C} \cup \mathcal{I}$.
   Output the relative keyword set $\mathcal{O}$.

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Example

Attribute	Tag	Keyword set
Title	Title	$\mathcal{C}_1$
Keyword	Keyword	$\mathcal{C}_2$
Author (as a related attribute)	Title $\Rightarrow$ Author $\Rightarrow$ Title	$\mathcal{E}$
Abstract	Abstract	$\mathcal{D}$

Association Rules Using Several Attributes (*Cont.*)

- Expanding keywords spaces (cont.)
  - Example (cont.)

Table 6. Derived keyword sets from keyword "bibliographic"

Keyword set	Keywords
$\mathcal{O}$	system, information, library, database, record, access, online, retrieval, service, CD, data
$\mathcal{C}_1$	database, information, system, library, record, online, retrieval, service
$\mathcal{C}_2$	database, information, library, system, online, record, retrieval, service, CD, data
$\mathcal{E}$	retrieval, system, hypertext, based, information, library, structure, linkage, semantics, searcher, rural, response, resources, performance, model, issue, effect, education, design, access, united, $\cdots$
$\mathcal{D}$	library, information, database, system, record, data, access, CD

Performances

- Comparison between full text search and the Weighted mining algorithm

Table 5. Performance of computing cost for our system.

Order	Normal [sec]	Total [sec]	Total/Normal [times]	Algorithm 1 [sec]	Algorithm 1 [%]	Algorithm 2 [sec]	Algorithm 2 [%]
1 - 100	27.8	54.7	2.0	17.6	27.9	18.4	20.3
101 - 200	27.5	56.5	2.1	17.7	28.0	18.9	20.1
201 - 300	27.3	58.5	2.1	19.4	29.7	18.8	19.2
301 - 400	26.8	63.5	2.4	25.1	34.6	18.3	17.6
401 - 500	26.6	66.6	2.5	27.0	36.1	18.5	17.0
501 - 600	27.2	65.3	2.4	24.2	32.5	18.9	18.1
601 - 700	26.6	69.2	2.6	28.4	36.0	18.6	16.6
701 - 800	27.0	67.8	2.5	26.6	34.2	18.6	17.0
801 - 900	26.8	69.1	2.6	28.0	35.9	18.6	16.7
901 - 1000	26.4	71.4	2.7	31.2	38.6	18.7	16.2

Performances (Cont.)

- Solutions
  - Sampling
    - reduce the deriving to pick up appropriate number of items
  - Caching
    - store query results for future use
  - Parallel processing