Applying Latent Semantic Indexing in Frequent Itemset Mining for Document Relation Discovery

Thanaruk Theeramunkong[1],
Kritsada Sriphaew[1,2](presenter)
and Manabu Okumura[2]

[1]School of Information and Computer Technology,
Sirindhorn International Institute of Technology, THAILAND

[2]Precision and Intelligence Laboratory,
Tokyo Institute of Technology, JAPAN

Outline

- What is our definition of document relation?
- Why we introduce LSI?
- How to evaluate the discovered relations?
- Results and Summary

Document Relation Definition

- In document network/graph area, a relation is introduced by edge or path between document nodes where an edge is introduced by hyperlink, author, citation, etc. [Kessler63, Garfield72, Small73, Chen99, An04]
- In IR and TM area, a relation is introduced by cosine similarity between query and the document vector or among two document vectors. respectively [Page98, Lawrence98, Baeza-Yates99, Theeramunkong04]

Document Relation Definition

- In document network/graph area, a relation is introduced by edge or path between document nodes where an edge is introduced by hyperlink, author, citation, etc. [Kessler63, Garfield72, Small73, Chen99, An04]
- In IR and TM area, a relation is introduced by cosine similarity between query and the document vector or among two document vectors. respectively [Page98, Lawrence98, Baeza-Yates99, Theeramunkong04]
- Usually, a relation is binary since it is introduced among only two documents
- **Our approach:** find document relations in which each relation introduce on n documents where n >= 2

Document Relation Definition

498 Medical News in Brief

Medical News in brief

CURRENT STATUS OF THERAPY
IN RHEUMATIC FEVER

In a report to the Council on Drugs of the A.M.A., McEwen (*J. A. M. A.*, 170: 1056, 1959) states that prevention of rheumatic fever is best achieved by prevention of hæmolytic streptococcic infections or by prompt intensive treatment of an existing infection. Continuous prophylaxis is carried out with benzathine penicillin G (one intramuscular injection of 1,200,000 units every four weeks), with sulfadiazine or a multiple sulfonamide in doses of 500 mg. or 1 gram daily in a single dose or penicillin tablets (buffered soluble

46

Clinical Medicine

Ethnic Differences in Mortality From Acute Rheumatic Fever and Chronic Rheumatic Heart Disease in New Mexico, 1958-1982

THOMAS M. BECKER, MD; CHARLES L. WIGGINS, MSPH; CHARLES R. KEY, MD; *and* JONATHAN M. SAMET, MD, *Albuquerque*

To examine time trends and differences in mortality rates from acute rheumatic fever and chronic rheumatic heart disease in New Mexico's Hispanic, American Indian, and non-Hispanic white populations, we analyzed vital records data for 1958 through 1982. Age-adjusted mortality rates for acute rheumatic fever were low and showed no consistent temporal trends among the three ethnic groups over the study period. Age-adjusted and age-specific mortality rates for chronic rheumatic heart disease in Hispanic and non-Hispanic whites decreased over the 25-year period, although rates were higher among Hispanics than among non-Hispanics during most of the time period. In American Indians, age-adjusted mortality rates for chronic rheumatic heart disease increased between 1968 and 1977 to twice the non-Indian mortality rates during the same period. Despite this increase in mortality from chronic rheumatic heart disease among New Mexico's American Indians from 1968 to 1977, the New Mexico data generally reflect national trends of decreasing mortality from chronic rheumatic heart disease.

(Becker TM, Wiggins CL, Key CR, et al: Ethnic differences in mortality from acute rheumatic fever and chronic rheumatic heart disease in New Mexico, 1958-1982. West J Med 1989 Jan; 150:46-50)

NEW MEXICO CARDIOVASCULAR DISEASE NURSING PROJECT

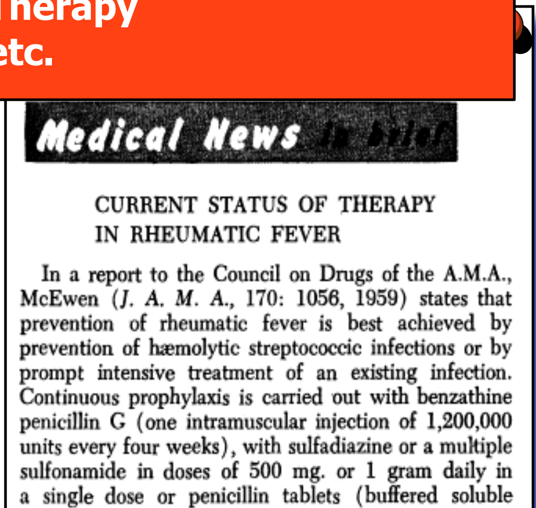
Stanley J. Leland, M.D., F.A.P.H.A.; Beatrice Chauvenet, M.A.; and Velma G. Long, B.S., R.N.

An extensive nursing education program in cardiovascular disease has been undertaken in New Mexico by the State Health Department in cooperation with public and voluntary agencies. Committee members from the New Mexico Heart Association, New Mexico Nurses' Association, the state's two

of the CVD Nursing Consultant in organizing and coordinating programs to meet their local needs. Ten community workshops in cardiovascular disease were held in the state during 1959, with a total attendance of 599 of whom 455 were registered nurses and 34 were physicians.

Document Relation Definition

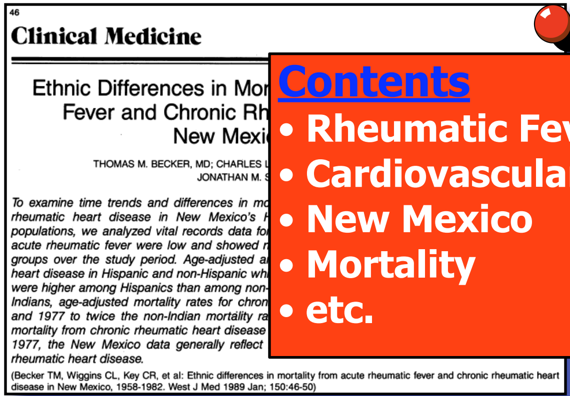
Contents
- Rheumatic Fever
- Cardiovascular disease
- Therapy
- etc.



Medical News in brief

CURRENT STATUS OF THERAPY IN RHEUMATIC FEVER

In a report to the Council on Drugs of the A.M.A., McEwen (*J. A. M. A.*, 170: 1056, 1959) states that prevention of rheumatic fever is best achieved by prevention of hæmolytic streptococcic infections or by prompt intensive treatment of an existing infection. Continuous prophylaxis is carried out with benzathine penicillin G (one intramuscular injection of 1,200,000 units every four weeks), with sulfadiazine or a multiple sulfonamide in doses of 500 mg. or 1 gram daily in a single dose or penicillin tablets (buffered soluble



46

Clinical Medicine

Ethnic Differences in Mor
Fever and Chronic Rh
New Mexi

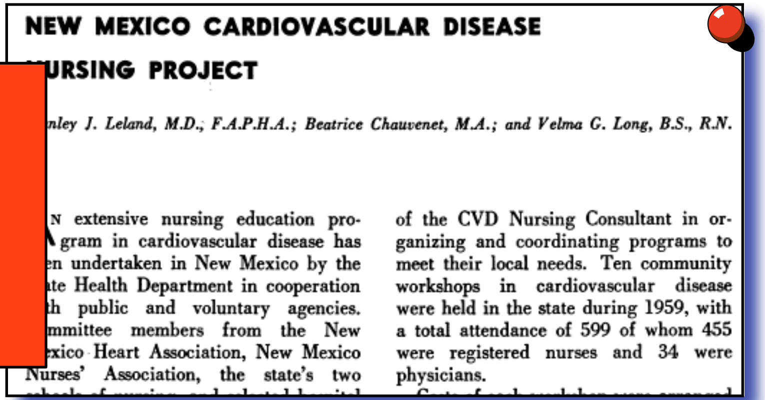
THOMAS M. BECKER, MD; CHARLES L
JONATHAN M. S

To examine time trends and differences in mo rheumatic heart disease in New Mexico's H populations, we analyzed vital records data fo acute rheumatic fever were low and showed r groups over the study period. Age-adjusted a heart disease in Hispanic and non-Hispanic wh were higher among Hispanics than among non- Indians, age-adjusted mortality rates for chron and 1977 to twice the non-Indian mortality ra mortality from chronic rheumatic heart disease 1977, the New Mexico data generally reflect rheumatic heart disease.

(Becker TM, Wiggins CL, Key CR, et al: Ethnic differences in mortality from acute rheumatic fever and chronic rheumatic heart disease in New Mexico, 1958-1982. West J Med 1989 Jan; 150:46-50)

Contents
- Rheumatic Fever
- Cardiovascular disease
- New Mexico
- Mortality
- etc.

Contents
- Cardiovascular disease
- New Mexico
- Nursing Project
- Therapy
- etc.



NEW MEXICO CARDIOVASCULAR DISEASE
URSING PROJECT

nley J. Leland, M.D.; F.A.P.H.A.; Beatrice Chauvenet, M.A.; and Velma G. Long, B.S., R.N.

N extensive nursing education pro- gram in cardiovascular disease has en undertaken in New Mexico by the ate Health Department in cooperation th public and voluntary agencies. mmittee members from the New xico Heart Association, New Mexico Nurses' Association, the state's two

of the CVD Nursing Consultant in or- ganizing and coordinating programs to meet their local needs. Ten community workshops in cardiovascular disease were held in the state during 1959, with a total attendance of 599 of whom 455 were registered nurses and 34 were physicians.

東京工業大学 精密工学研究所 Precision and Intelligence Laboratory

SIIT

Document Relation Definition

498 MEDICAL NEWS IN BRIEF

Medical News in brief

CURRENT STATUS OF THERAPY
IN RHEUMATIC FEVER

In a report to the Council on Drugs of the A.M.A., McEwen (*J. A. M. A.*, 170: 1056, 1959) states that prevention of rheumatic fever is best achieved by prevention of hæmolytic streptococcic infections or by prompt intensive treatment of an existing infection. Continuous prophylaxis is carried out with benzathine penicillin G (one intramuscular injection of 1,200,000 units every four weeks), with sulfadiazine or a multiple sulfonamide in doses of 500 mg. or 1 gram daily in a single dose or penicillin tablets (buffered soluble

Related Contents
- **Rheumatic Fever**
- **Cardiovascular disease**

46

Clinical Medicine

Ethnic Differences in Mortality From Acute Rheumatic Fever and Chronic Rheumatic Heart Disease in New Mexico, 1958-1982

...AS M. BECKER, MD; CHARLES L. WIGGINS, MSPH; CHARLES R. KEY, MD; and JONATHAN M. SAMET, MD, *Albuquerque*

To examine time trends and differences in mortality rates from acute rheumatic fever and chronic rheumatic heart disease in New Mexico's Hispanic, American Indian, and non-Hispanic white populations, we analyzed vital records data for 1958 through 1982. Age-adjusted mortality rates for acute rheumatic fever were low and showed no consistent temporal trends among the three ethnic groups over the study period. Age-adjusted and age-specific mortality rates for chronic rheumatic heart disease in Hispanic and non-Hispanic whites decreased over the 25-year period, although rates were higher among Hispanics than among non-Hispanics during most of the time period. In American Indians, age-adjusted mortality rates for chronic rheumatic heart disease increased between 1968 and 1977 to twice the non-Indian mortality rates during the same period. Despite this increase in mortality from chronic rheumatic heart disease among New Mexico's American Indians from 1968 to 1977, the New Mexico data generally reflect national trends of decreasing mortality from chronic rheumatic heart disease.

(Becker TM, Wiggins CL, Key CR, et al: Ethnic differences in mortality from acute rheumatic fever and chronic rheumatic heart disease in New Mexico, 1958-1982. West J Med 1989 Jan; 150:46-50)

Related Contents
- **Cardiovascular disease**
- **New Mexico**

NEW MEXICO CARDIOVASCULAR DISEASE
NURSING PROJECT

Stanley J. Leland, M.D., F.A.P.H.A.; Beatrice Chauvenet, M.A.; and Velma G. Long, B.S., R.N.

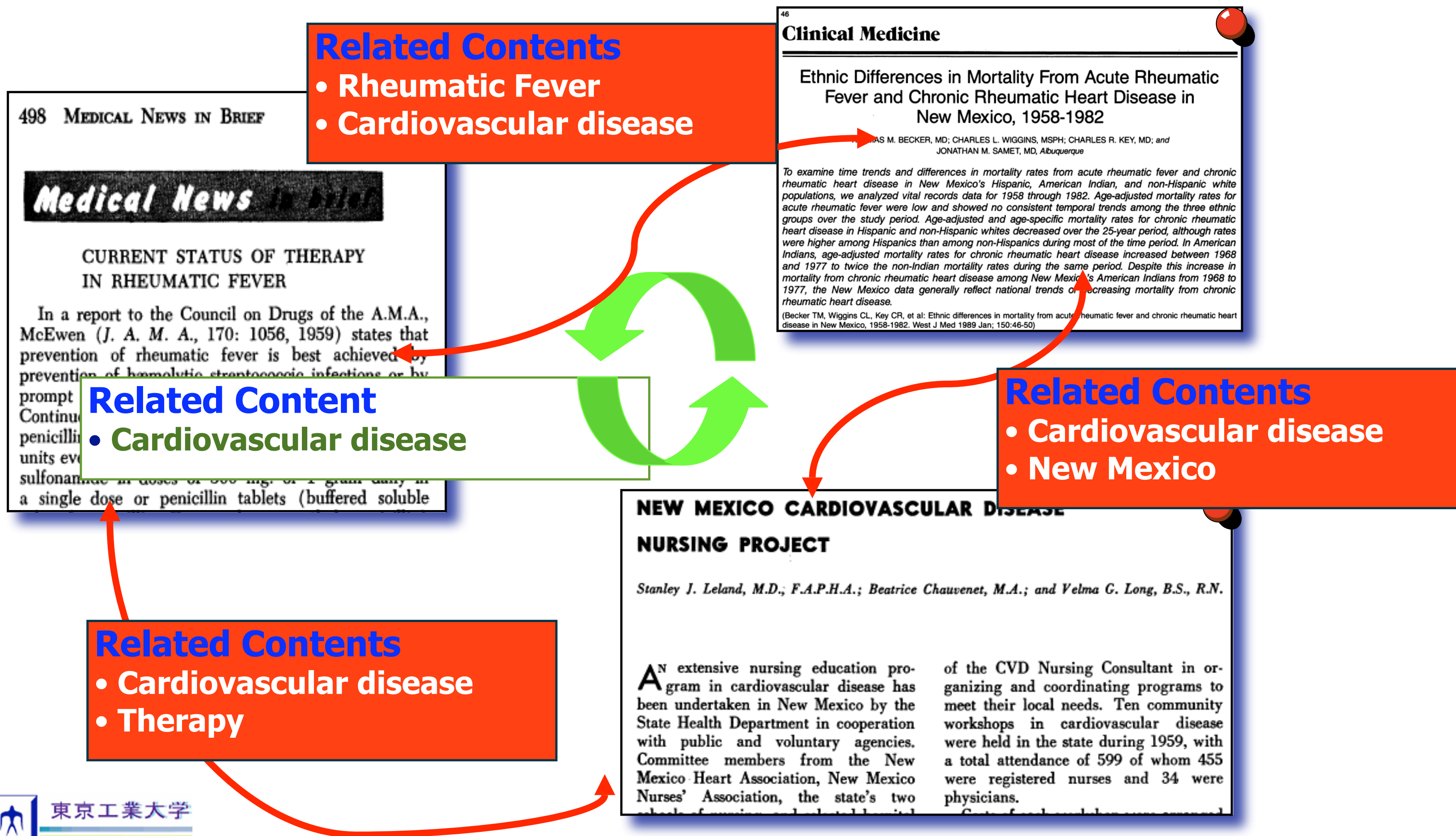
An extensive nursing education program in cardiovascular disease has been undertaken in New Mexico by the State Health Department in cooperation with public and voluntary agencies. Committee members from the New Mexico Heart Association, New Mexico Nurses' Association, the state's two of the CVD Nursing Consultant in organizing and coordinating programs to meet their local needs. Ten community workshops in cardiovascular disease were held in the state during 1959, with a total attendance of 599 of whom 455 were registered nurses and 34 were physicians.

Related Contents
- **Cardiovascular disease**
- **Therapy**

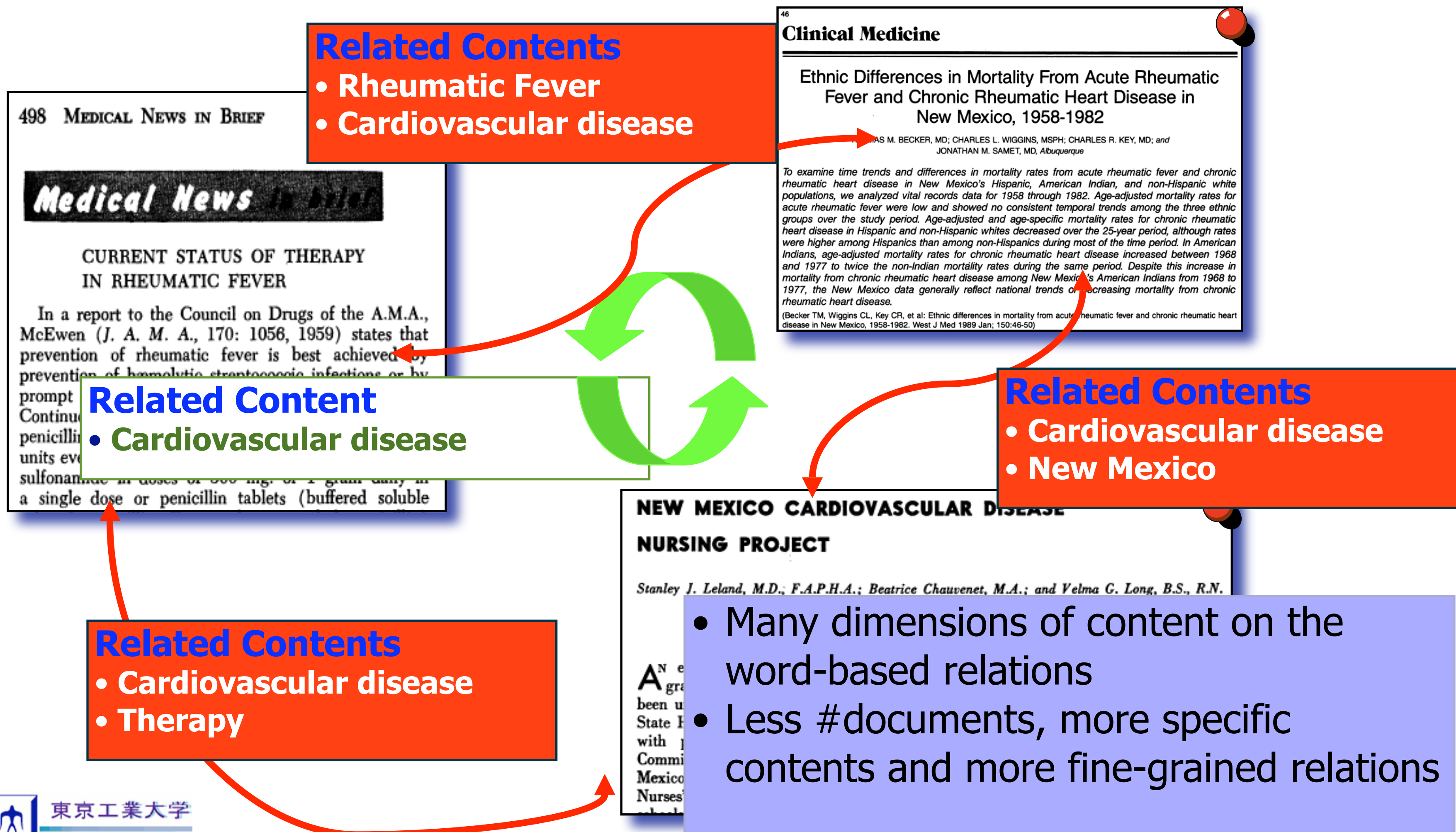
SIIT

東京工業大学
精密工学研究所
Precision and Intelligence Laboratory

Document Relation Definition



Document Relation Definition



- Many dimensions of content on the word-based relations
- Less #documents, more specific contents and more fine-grained relations

Method for document relation discovery

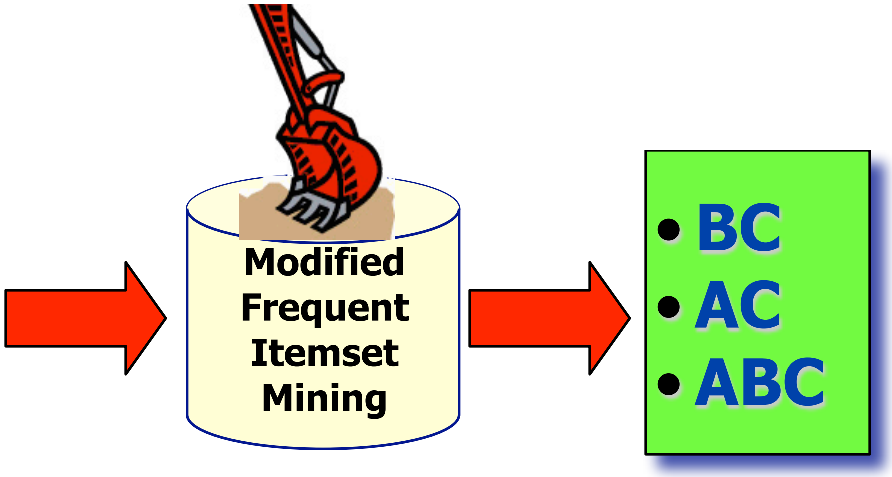
Modified Frequent Itemset Mining [Sriphaew05, Sriphaew07]

Term	document			
	A	B	C	D
Data	4	1	4	2
Mining	2	5	3	2
Association	0	3	1	1
Rule	0	4	1	1
Technique	2	1	2	1
Data Mining	2	0	1	1
Association Rule	0	4	1	1

Method for document relation discovery

Modified Frequent Itemset Mining [Sriphaew05, Sriphaew07]

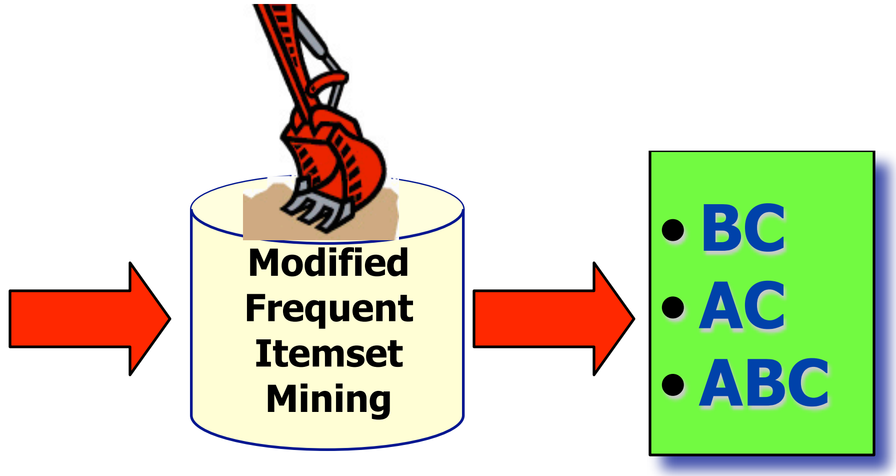
Term	document			
	A	B	C	D
Data	4	1	4	2
Mining	2	5	3	2
Association	0	3	1	1
Rule	0	4	1	1
Technique	2	1	2	1
Data Mining	2	0	1	1
Association Rule	0	4	1	1



Method for document relation discovery

Modified Frequent Itemset Mining [Sriphaew05, Sriphaew07]

Term	document A	document B	document C	document D
Data	4	1	4	2
Mining	2	5	3	2
Association	0	3	1	1
Rule	0	4	1	1
Technique	2	1	2	1
Data Mining	2	0	1	1
Association Rule	0	4	1	1



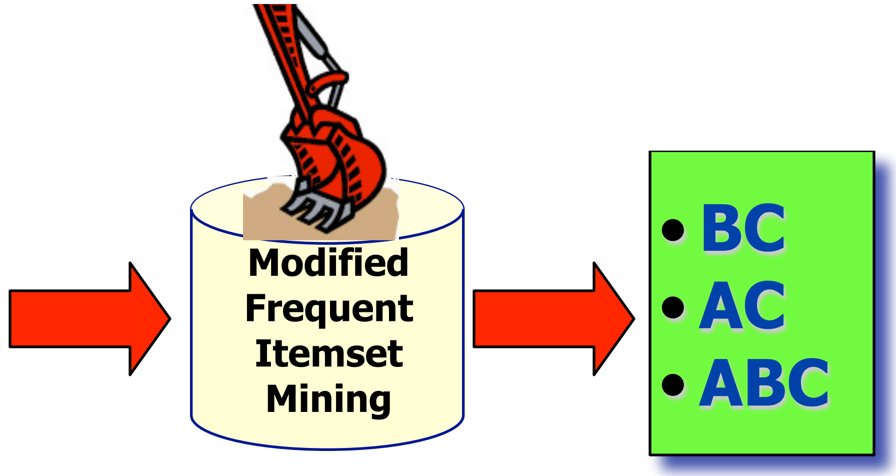
$$sup(X_k) = \frac{\sum_{j=1}^{n} min_{i=1}^{k} w(x_i,t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} w(d_i,t_j)}$$

- $w(x_i,t_j)$ is a weight of term t_j in document x_i
- preserves closure properties

Method for document relation discovery

Modified Frequent Itemset Mining [Sriphaew05, Sriphaew07]

Term	document			
	A	B	C	D
Data	4	1	4	2
Mining	2	5	3	2
Association	0	3	1	1
Rule	0	4	1	1
Technique	2	1	2	1
Data Mining	2	0	1	1
Association Rule	0	4	1	1



$$sup(X_k) = \frac{\sum_{j=1}^{n} min_{i=1}^{k} w(x_i,t_j)}{\sum_{j=1}^{n} max_{i=1}^{m} w(d_i,t_j)}$$

- $w(x_i,t_j)$ is a weight of term t_j in document x_i
- preserves closure properties

<u>Why FIM?</u>

– Possible to apply cosine similarity on every combination of document sets
– But our target is set of documents (involve more than two documents)
– Pruning strategy exists
– Several efficient algorithms

Problems of Document Representation

- Existing approach directly exploits words/terms in documents to discover relations using word co-occurrences and shared vocabularies.
- A relation on a set of documents may occur even if they do not share any common words or terms but their terms are semantically related.

Why we use LSI?

- We want some terms which are semantically related to existing terms in a document to have some weights while reducing meaningless terms (terms appear in small eigen vector)

- We still want to have term-document matrix where we can apply FIM to discover document relations, therefore, PCA which its input is covariance matrix is not our case.

Latent Semantic Indexing (LSI)

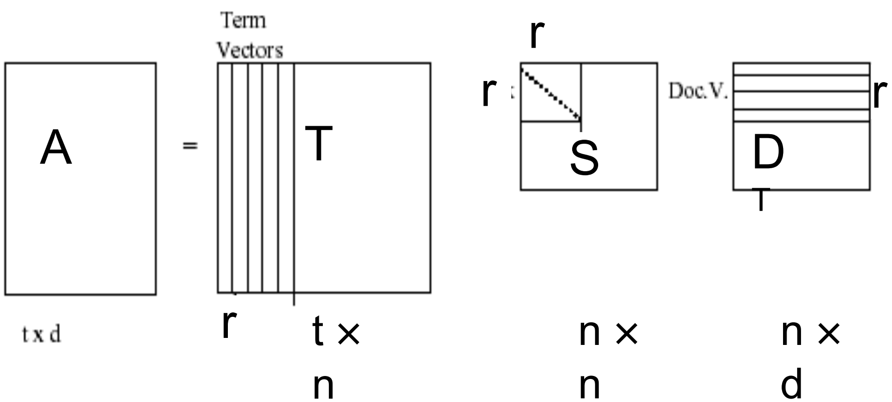
$$A_{t\times d} = T_{t\times n} \times S_{n\times n} \times D_{d\times n}{}^{T}$$

- A is the input document-term matrix
- t is the number of terms
- d is the number of documents
- $n = \min(t,d)$
- T and D have orthonormal columns $T\times T^T=I$ and $D^T\times D = I$
- S is a diagonal matrix where $s_{i,j} = 0$ for $i \neq j$

LSI uses an SVD method to decompose the input A and represents it as A′ with the objective function: $\min \| A - A' \|_2$

Latent Semantic Indexing (LSI)

$$A_{t\times d} = T_{t\times n} \times S_{n\times n} \times D_{d\times n}^{T}$$



- A is the input document-term matrix
- t is the number of terms
- d is the number of documents
- $n = \min(t,d)$
- T and D have orthonormal columns $T\times T^{T}=I$ and $D^{T}\times D = I$
- S is a diagonal matrix where $s_{i,j} = 0$ for $i \neq j$

LSI uses an SVD method to decompose the input A and represents it as A′ with the objective function: **$\min \| A - A' \|_2$**

In some cases, rank(A) = r where $r \leq n$, the diagonal elements of S are $\sigma_1, \sigma_2, \ldots, \sigma_n$ *where* $\sigma_i > 0$ *for* $1 \leq i \leq r$ and $\sigma_i = 0$ *for* $r < i \leq n$

In this work, we get potential σ_i by using simple kappa statistics, i.e., relative accuracy of Ritz values acceptable as eigenvalue >= 1.00E-06

Proposed threshold δ

- Since some meaningless terms will have some wieghts after dimension reduction, therefore, we want to filter out those terms.
- We generate new document-term matrix A″ where,

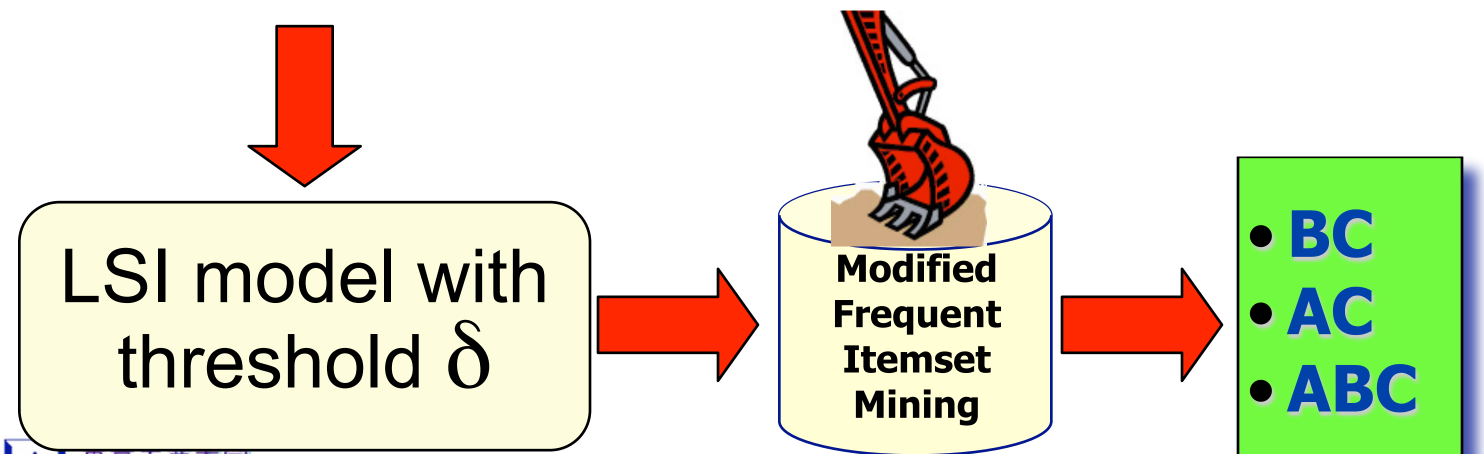
$$A''_{txd} = [a''_{ij}],\ 1 \le i \le t \ and\ 1 \le j \le d$$

$$a''_{ij} = \begin{cases} a'_{ij}, if\ a'_{ij} \ge \delta \\ 0 \quad , otherwise \end{cases}$$

Objectives

Term	document			
	A	B	C	D
Data	4	1	4	2
Mining	2	5	3	2
Association	0	3	1	1
Rule	0	4	1	1
Technique	2	1	2	1
Data Mining	2	0	1	1
Association Rule	0	4	1	1

- Study effects of different weighting, **tf** and **tf-idf** for new modified approach of FIM
- Study effects of LSI with δ threshold and the quality of document relations



Evaluation Concept

Problems:

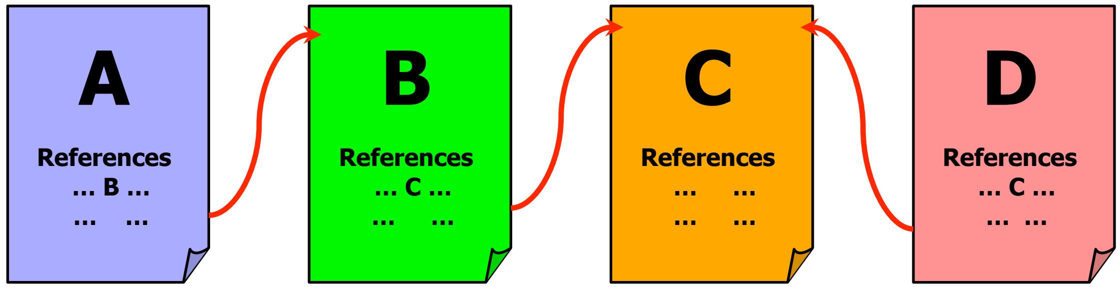
- Lack of corpus with correct answers
- Excessive time-consuming and labor-intensive task for human evaluation

  For example, we need to investigate ${}^{10000}C_2 \approx 50\times10^6$ pairs if we want to construct a corpus with 10,000 documents

Solution Idea:

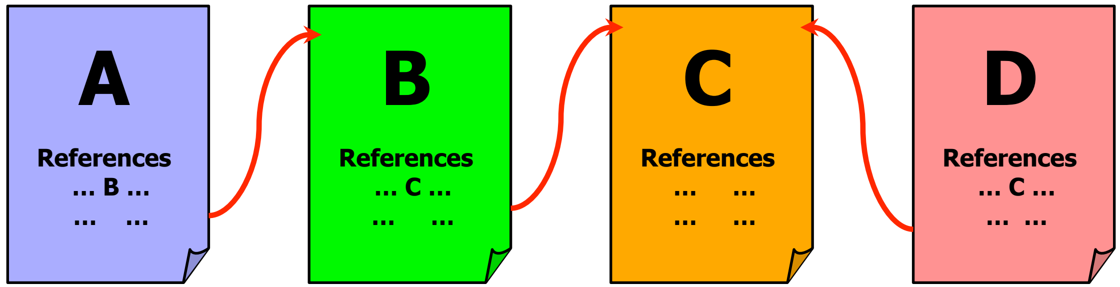
- Use other potential relation information as comparative criteria
- Trust knowledge for evaluation **citations** (or references) in research articles.
- **Remark:** our approach discovers word-based relations
  but we make comparison with the citation-based relations w/o using
  citation information for discovery

Evaluation Concept



- Formulating the evaluation criteria as an "**Ordered Accumulative Citation Matrix**" (**OACM**) using the citation information and the transitivity function

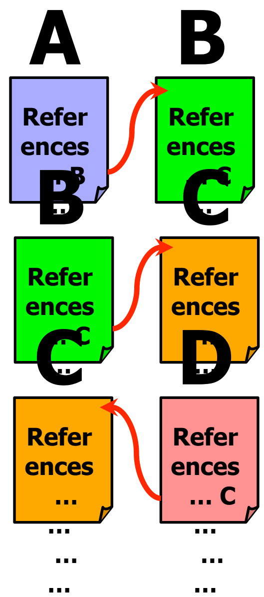
Evaluation Concept



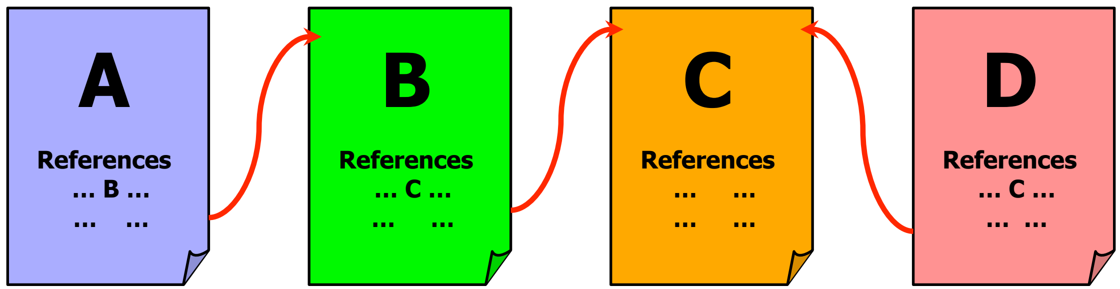
- First Criteria:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right) \end{array}$$

1-OACM



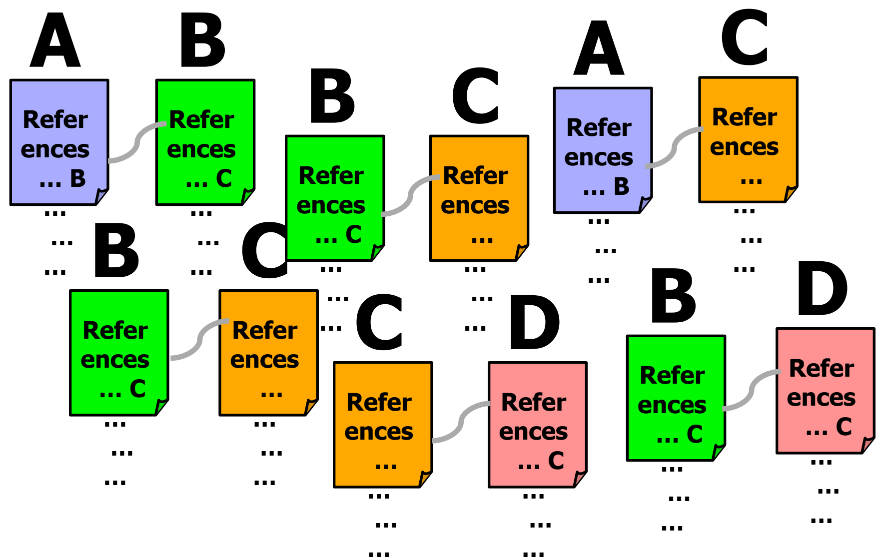
Evaluation Concept



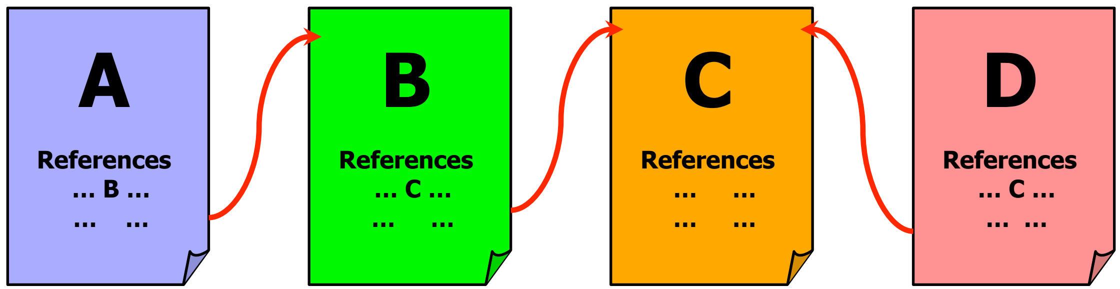
- Second Criteria:

$$
\begin{array}{c|cccc|}
 & A & B & C & D \\
A & 1 & 1 & 1 & 0 \\
B & 1 & 1 & 1 & 1 \\
C & 1 & 1 & 1 & 1 \\
D & 0 & 1 & 1 & 1
\end{array}
$$

2-OACM



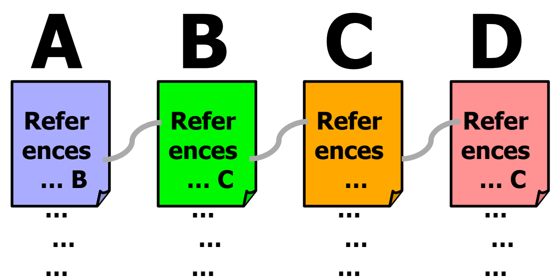
Evaluation Concept



- Third Criteria:

$$
\begin{array}{c|cccc}
 & A & B & C & D \\
\hline
A & 1 & 1 & 1 & 1 \\
B & 1 & 1 & 1 & 1 \\
C & 1 & 1 & 1 & 1 \\
D & 1 & 1 & 1 & 1
\end{array}
$$

3-OACM



We stop at third criteria (3-OACM) since there is no significant difference after third criteria in our preliminary experiments

Evaluation Concepts

- Proposed Scoring method: Counting the valid relations based on evaluation criteria

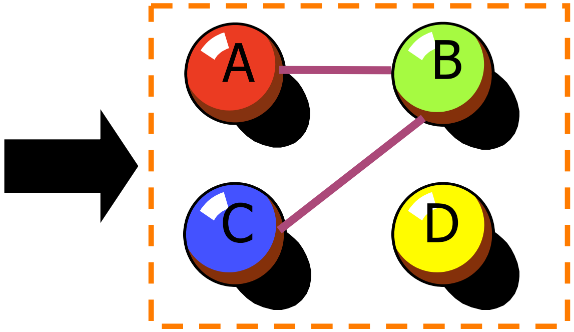
Evaluation Concepts

- Proposed Scoring method: Counting the valid relations based on evaluation criteria

Evaluation Criteria: <u>**1-OACM**</u>

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 1 & 1 & 0 & 0 \\ B & 1 & 1 & 1 & 0 \\ C & 0 & 1 & 1 & 1 \\ D & 0 & 0 & 1 & 1 \end{array}$$

Based on 1-OACM



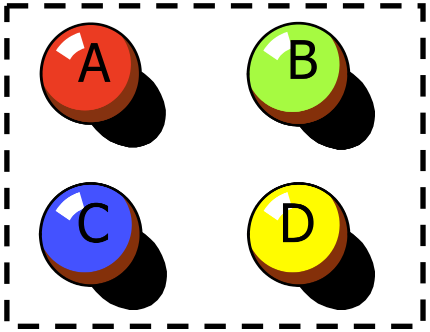
Evaluation Concepts

- Proposed Scoring method: Counting the valid relations based on evaluation criteria

Discovered document relations

Evaluation Criteria: **1-OACM**

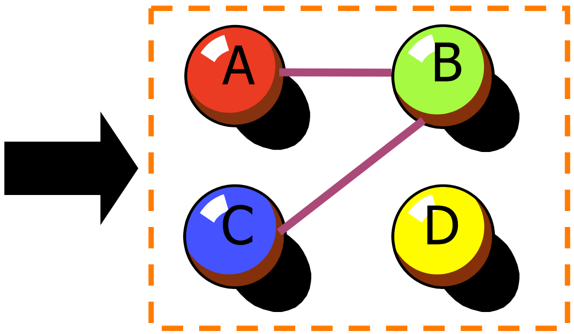
$$\begin{array}{c c} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$



ABCD

Validity of discovered relation ABCD based on 1-OACM = 2/3 = 0.67

Based on 1-OACM



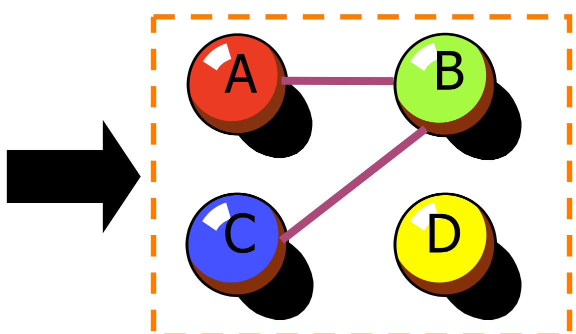
Evaluation Concepts

- Proposed Scoring method: Counting the valid relations based on evaluation criteria

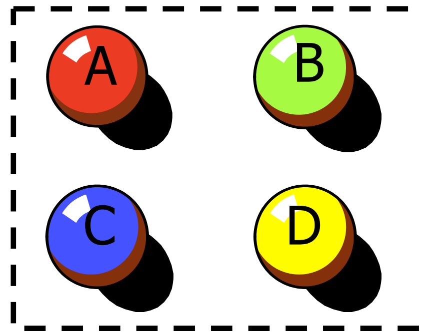
Evaluation Criteria: **1-OACM**

$$
\begin{array}{c}
 \\ A \\ B \\ C \\ D
\end{array}
\begin{array}{c}
\begin{array}{cccc} A & B & C & D \end{array} \\
\left(\begin{array}{cccc}
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1
\end{array}\right)
\end{array}
$$

Based on 1-OACM



Discovered document relations



ABCD

Validity of discovered relation ABCD based on 1-OACM = 2/3 = 0.67

For all discovered set,

we use **weighted mean** of validity as evaluation measurement where the weight is given by the number of documents in each discovered relation.

Dataset

Test Collection

- **10,817** scientific research articles*
- 3 classes: Hardware, Data, Computer
- Extract citation network to form evaluation criteria but exclude those texts from data
- Preprocessing: filtering stopwords, terms occur <3 times and bigram

* Articles are collected from ACM Digital Library

Experimental Results

N is Top-N rankings (by support) of discovered relations when either tf/tf-idf is used and LSI is applied with different δ thresholds

Methods	N	1-OACM		2-OACM		3-OACM	
		tf	tfidf	tf	tfidf	tf	tfidf
w/o LSI	1000	14.29	25.00	85.71	100.00	100.00	100.00
	5000	37.59	38.03	87.23	95.77	95.62	97.18
	10000	18.22	38.97	58.94	87.66	87.13	93.81
	50000	6.16	16.24	35.91	60.52	75.68	94.05
	100000	4.37	14.36	31.22	55.83	74.49	93.08
$LSI_{\delta=0.5}$	1000	41.51	42.86	90.57	85.71	94.34	91.43
	5000	23.80	25.90	66.47	67.94	84.01	83.76
	10000	19.92	23.01	64.44	67.26	86.06	85.02
	50000	14.12	17.89	59.80	64.13	90.15	89.13
	100000	11.40	14.48	56.81	60.57	90.39	90.13

1. w/o LSI,
   tfidf is better than tf
   tfidf can help to find relations for direct use of words/terms
2. LSI case,
   Applying LSI is better than w/o LSI
   tf is better than tfidf is better than idf degrades the performance of LSI

Experimental Results

N is Top-N rankings (by support) of discovered relations when either tf/tf-idf is used and LSI is applied with different δ thresholds

Methods	N	1-OACM		2-OACM		3-OACM	
		tf	tfidf	tf	tfidf	tf	tfidf
w/o LSI	1000	14.29	25.00	85.71	100.00	100.00	100.00
	5000	37.59	38.03	87.23	95.77	95.62	97.18
	10000	18.22	38.97	58.94	87.66	87.13	93.81
	50000	6.16	16.24	35.91	60.52	75.68	94.05
	100000	4.37	14.36	31.22	55.83	74.49	93.08
$LSI_{\delta=0.5}$	1000	41.51	42.86	90.57	85.71	94.34	91.43
	5000	23.80	25.90	66.47	67.94	84.01	83.76
	10000	19.92	23.01	64.44	67.26	86.06	85.02
	50000	14.12	17.89	59.80	64.13	90.15	89.13
	100000	11.40	14.48	56.81	60.57	90.39	90.13
$LSI_{\delta=0.7}$	1000	47.14	44.15	90.00	80.32	95.71	85.64
	5000	25.95	28.28	69.09	70.86	85.98	85.72
	10000	22.26	25.59	67.80	70.64	87.52	86.95
	50000	14.77	19.91	60.76	66.72	91.43	91.27
	100000	12.09	16.06	57.51	61.73	91.52	90.98
$LSI_{\delta=1.0}$	1000	44.68	45.42	85.11	81.25	90.43	87.08
	5000	26.55	28.95	70.23	71.42	86.86	86.43
	10000	23.67	27.85	69.27	72.66	88.54	89.15
	50000	15.27	19.79	61.05	66.58	91.75	91.29
	100000	12.53	16.45	57.35	62.03	91.67	91.90

Experimental Results

N is Top-N rankings (by support) of discovered relations when either tf/tf-idf is used and LSI is applied with different δ thresholds

Methods	N	1-OACM		2-OACM		3-OACM	
		tf	tfidf	tf	tfidf	tf	tfidf
w/o LSI	1000	14.29	25.00	85.71	100.00	100.00	100.00
	5000	37.59	38.03	87.23	95.77	95.62	97.18
	10000	18.22	38.97	58.94	87.66	87.13	93.81
	50000	6.16	16.24	35.91	60.52	75.68	94.05
	100000	4.37	14.36	31.22	55.83	74.49	93.08
$LSI_{\delta=0.5}$	1000	41.51	42.86	90.57	85.71	94.34	91.43
	5000	23.80	25.90	66.47	67.94	84.01	83.76
	10000	19.92	23.01	64.44	67.26	86.06	85.02
	50000	14.12	17.89	59.80	64.13	90.15	89.13
	100000	11.40	14.48	56.81	60.57	90.39	90.13
$LSI_{\delta=0.7}$	1000	47.14	44.15	90.00	80.32	95.71	85.64
	5000	25.95	28.28	69.09	70.86	85.98	85.72
	10000	22.26	25.59	67.80	70.64	87.52	86.95
	50000	14.77	19.91	60.76	66.72	91.43	91.27
	100000	12.09	16.06	57.51	61.73	91.52	90.98
$LSI_{\delta=1.0}$	1000	44.68	45.42	85.11	81.25	90.43	87.08
	5000	26.55	28.95	70.23	71.42	86.86	86.43
	10000	23.67	27.85	69.27	72.66	88.54	89.15
	50000	15.27	19.79	61.05	66.58	91.75	91.29
	100000	12.53	16.45	57.35	62.03	91.67	91.90

1. δ threshold
   There is a suitable threshold to achieve highest validity
   LSI helps to discover direct citati
   more than indirect citations (we le
   method to
   optimize δ threshold as future work)

2. LSI case & OACM,
   1-OACM
   Higher δ is better than Lower δ
   2-,3-OACMs
   Lower δ is better than Higher δ
   LSI helps to discover direct citati
   more than indirect citations

Conclusions

- This work presents new approach to discover document relations using FIM and applying LSI for improving good document representation
- The quality of discovered document relations from our word-based approach can be relatively compared with ones from citation network. Those relations may be not the same kind of relations, but they shows good relation between those two kinds of document relations
- LSI is helpful to discover meaningful document relations especially the relations that is identical to **direct citations** whereas we still have indirect citations in the top ranks of discovered relations

Discussions and Future Works

- Some weak points of this work:
  - Testing on one corpus since it is difficult to construct large enough data of this kind.
  - Starting research for a new problem of document relation discovery where each relation composes of two or more documents. Therefore, there is no other method addressed the same problem with us. Although we can modify other existing methods for this task, we just want to sketch up the solution and fulfill all necessary processes for document relations especially the evaluation concept
- Low validity does not mean bad relations but it is not coincident with citation relations. Our method performs well in detect jeopardize articles and some novel relations which is not introduced by citations
- Exploring other term weighting and dimension reduction approaches

Thank you