Patent Document Retrieval and Classification at KAIST

KAIST CS Dept. / BOLA

2005. 12. 8.

Jae-Ho Kim, Jin-Xia Huang, Ha-Yong Jung, Key-Sun Choi

Introduction

- Tasks
  - Document retrieval subtask
  - Theme categorization subtask
- Our Approach
  - Patent retrieval: find most important parts
    - Patent document is structured
    - Re-organize patents according to detailed descriptions
  - Patent categorization: noisy elimination
    - K-NN based: eliminate un-important documents and un-important words in categorization
    - MEM based: only eliminate un-important words in categorization
    - *By using re-organized patents!*

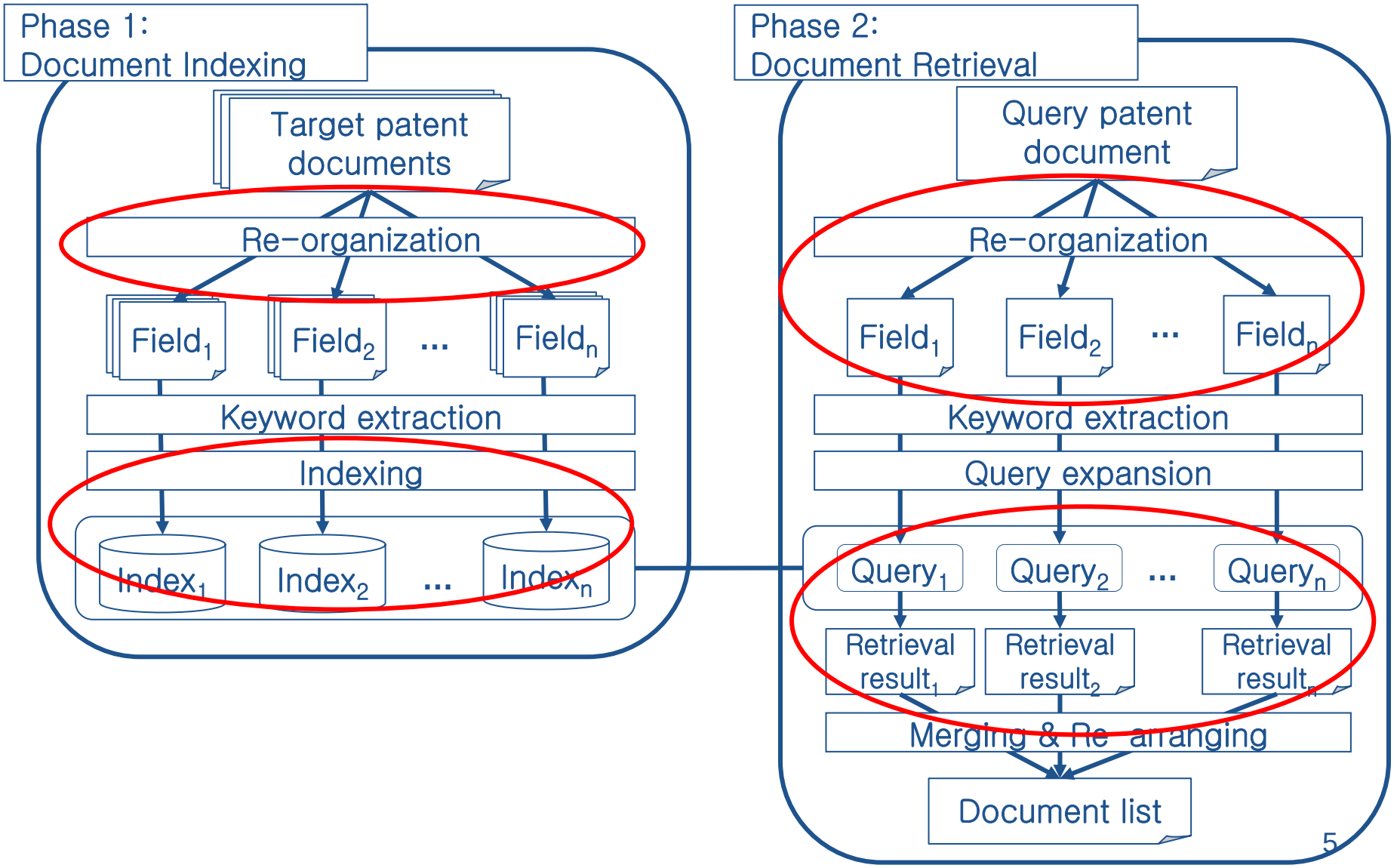
Patent Retrieval

- Finding the most important components of patent
- Perform IR by component-by-component comparison

Patent Retrieval

- Find most important parts
  - Patent document is structured
  - Re-organize patents according to detailed descriptions
- Component-by-component comparison
  - Two documents are similar, if the two documents are in the same technical classes and have the same (or similar) problem and solution (method).

Patent Retrieval: Find most important parts



Patent document is structured

Normative section

	<DOCNO>PATENT-JA-UPA-1995-000001<
<Bibliography> [publication date] [title of invention]	**<SDO BIJ>** (43)【公開日】平成7年（1995）1月6日 (54)【発明の名称】スラリ散布を行う土壌作業機 .....
<Abstract> [purpose] [composition]	**<SDO ABJ>** 【目的】スラリの処理と土壌作業を同時に行うことで、…… 【構成】トラクタとスラリを積載したバキウムカーとの間に ……
<Claims> [claim1] [claim2]	**<SDO CLJ>** 【請求項1】バキウムカーを牽引 【請求項2】トラクタに対して3点リ
<Description> [industrial application field] [problem to be solved] [means of solving problems] [operation] [embodiment examples] [effects of invention]	**<SDO DEJ>** 【産業上の利用分野】本発明はスラリ散布を行う土壌作業機に関し、…… 【発明が解決しようとする課題】このようなスラリを圃場に供給する…… 【課題を解決するための手段】上述のような目的を達成するために、…… 【作用】本発明のスラリ散布を行う土壌作業機は、…… 【実施例】以下、本発明を採用した土壌作業機について添付した図面に…… 【発明の効果】以上の説明から明らかなように、……
<Explanation of Drawings>	**<SDO EDJ>**【図1】本発明のスラリ散布を行う土壌作業機の側面図である。
<Drawings>	**<SDO DRJ>**【図1】

Detailed component

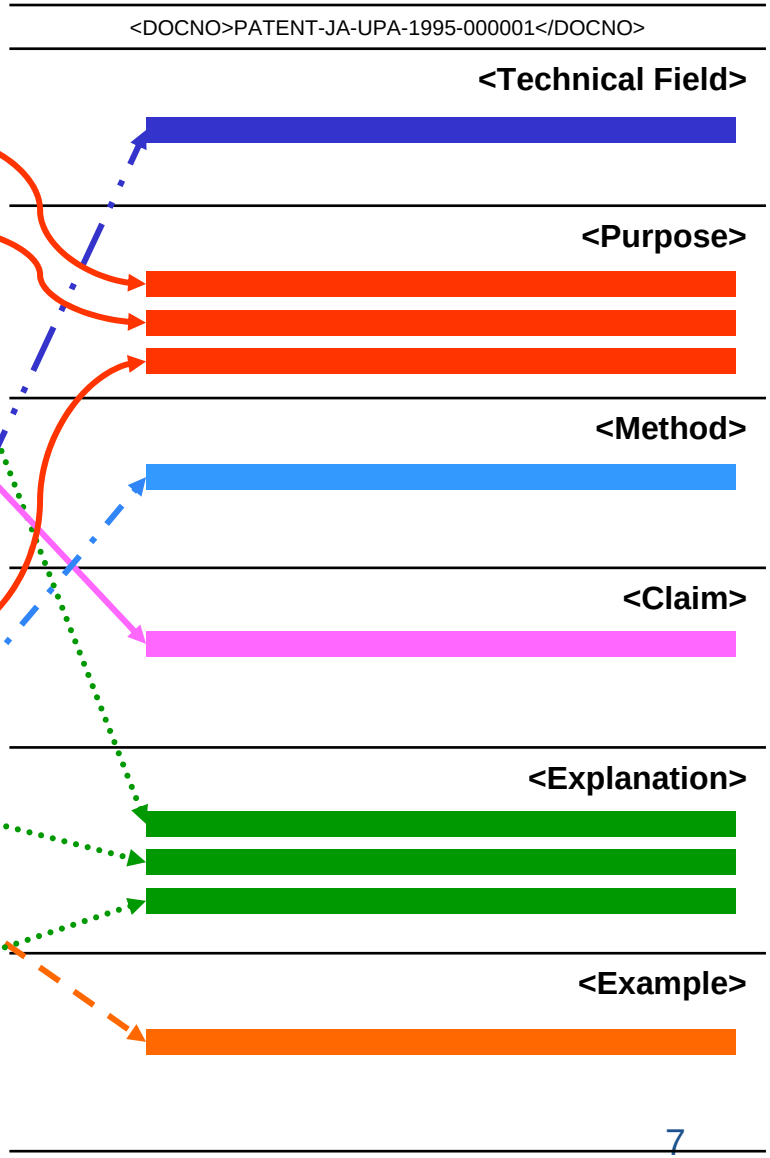
Applicant-defined tags

Re-organization of Patent Documents

<DOCNO>PATENT-JA-UPA-1995-000001</DOCNO>

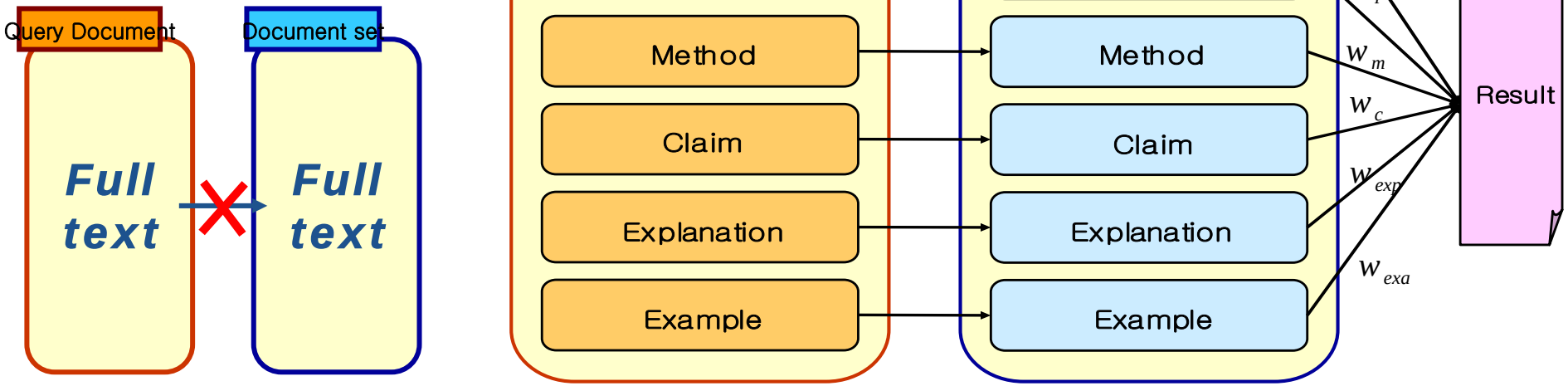
<SDO BIJ>	**<Bibliography>**
(43)【公開日】平成７年（１９９５）……	[Publication Date]
(54)【発明の名称】スラリ散布を行う……	[Title of Invention]
……	
<SDO ABJ>	**<Abstract>**
【目的】 スラリの処理と土壌作業を……	[Purpose]
【構成】 トラクタとスラリを積載した……	[Composition]
<SDO CLJ>	**<Scope of Claim>**
【請求項１】 バキウムカーを牽引して……	[Claim1]
【請求項２】 トラクタに対して ……	[Claim2]
<SDO DEJ>	**<Detailed Explanation>**
【産業上の利用分野】本発明はスラリ……	[Application Field]
【発明が解決しようとする課題】このようなスラリを圃場に供給する……	[Problem to be solved]
【課題を解決するための手段】上述のような目的を達成するために、……	[Means of solving Problems]
【作用】本発明のスラリ散布を行う土壌作業機は、……	[Operation]
【実施例】以下、本発明を採用した土壌作業機について添付した図面に……	[Embodiment Example]
【発明の効果】以上の説明から明らかな……	[Effects of Invention]
<SDO EDJ>	**<Explanation of Drawings>**
【図１】本発明のスラリ散布を行う.....	[Figure1]
<SDO DRJ>	**<Drawings>**
【図１】	[Figure1]
.....	

<DOCNO>PATENT-JA-UPA-1995-000001</DOCNO>



Find Most Important Parts in Patent Documents

- Patent Retrieval
  - Throw 6 queries to 6 indexes
    - The pairs of same fields are compared
    - Using keywords and BM 25 Okapi (by using Lemur Toolkit)
  - Use weighted summation for final result

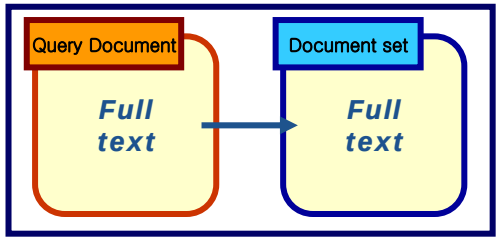


Experiment Results in Document Retrieval Subtask

- Baseline
  - Retrieval using full documents
- Post processing for increasing precision
  - Re-retrieval from retrieved 6,000 documents
    - by the comparison of Noun-Verb pairs

$$score = score_{original} + \beta \cdot score_{re-retrieval}$$

- *Using re-organized patents showed better results!*



RunId	Condition	Topics			
		a.ntc4	b.ntc4	a.ntc5	b.ntc5
baseline	= 0.00	0.1362	0.1286	0.1419	0.1181
d0010	β = 0.00	0.1576	0.1488	0.1642	0.1366
d0011	β = 0.10	0.1620	0.1473	0.1675	0.1396
d0012	β = 0.15	0.1655	0.1489	0.1647	0.1368
d0013	β = 0.20	0.1621	0.1453	0.1608	0.1334
d0014	β = 0.25	0.1608	0.1397	0.1591	0.1306
d0015	β = 0.30	0.1594	0.1381	0.1573	0.1286

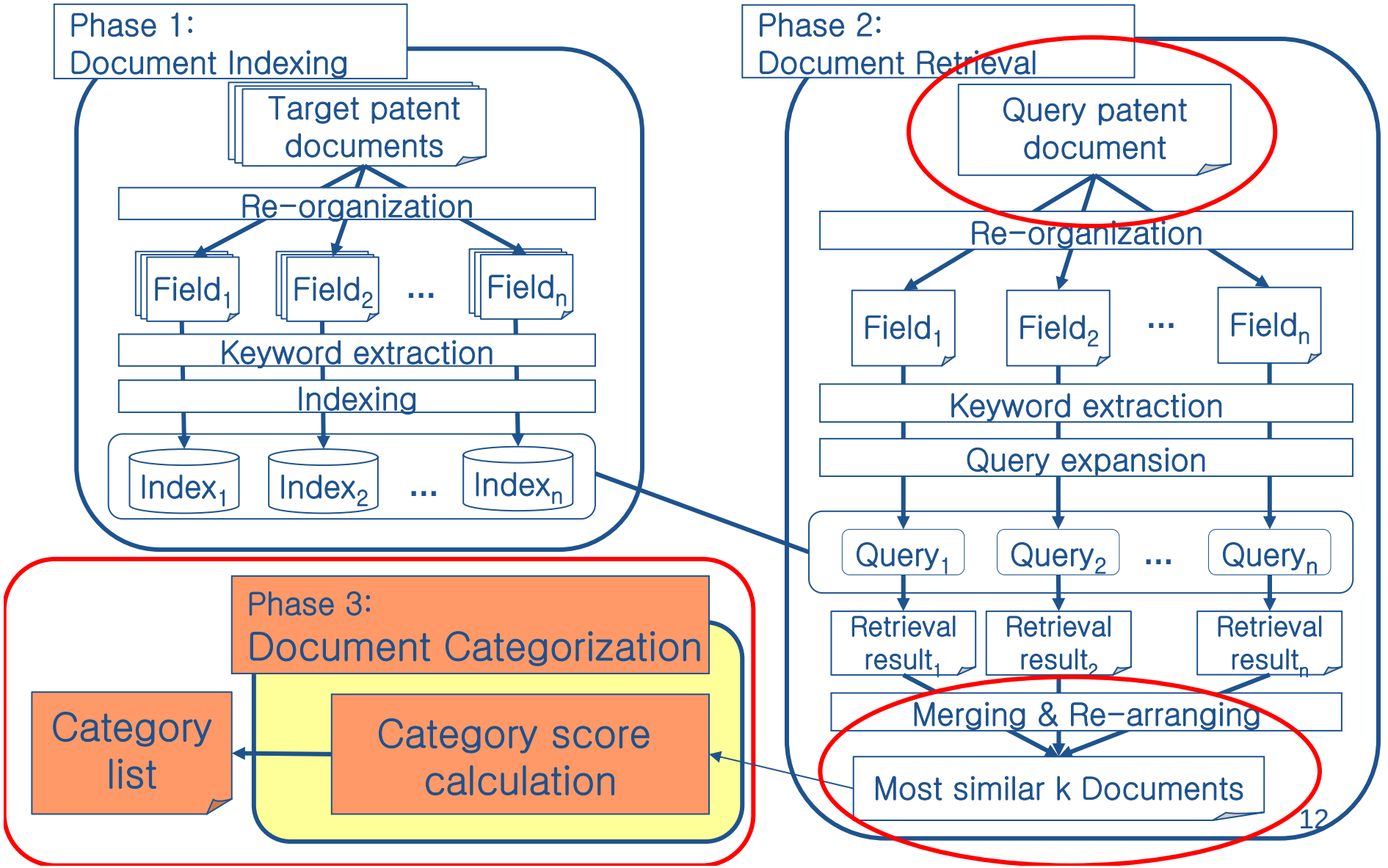
Patent Categorization

- Using the most important components of patent
- Perform categorization by k-NN based and MEM based approaches: different in noisy elimination

Patent Categorization

- Use only the most important components
  - Use re-organized patents, which has been proved helpful in patent retrieval
- Different noisy elimination approaches
  - K-NN based: classifies a given patent into the theme codes of k documents similar to it
    - Eliminate un-important documents by patent retrieval
    - Eliminate un-important words by keyword extraction
  - MEM based: classifies by similarity calculation between different document vectors
    - Eliminate un-important words by using weighing functions and threshold

kNN-based Patent Categorization



MEM-based Patent Categorization

- Feature selection to reduce the feature size
  - Vector: term vector with a word bag
  - Weighting function: TF/IDF and TF/ICF
    - CF(w): the category frequency of a word *w*

$$TFICF(w_i^{(d)}) = TF(w_i, d) \cdot ICF(w_i) \qquad ICF(w_i) = \log(\frac{|C|}{CF(w_i)})$$

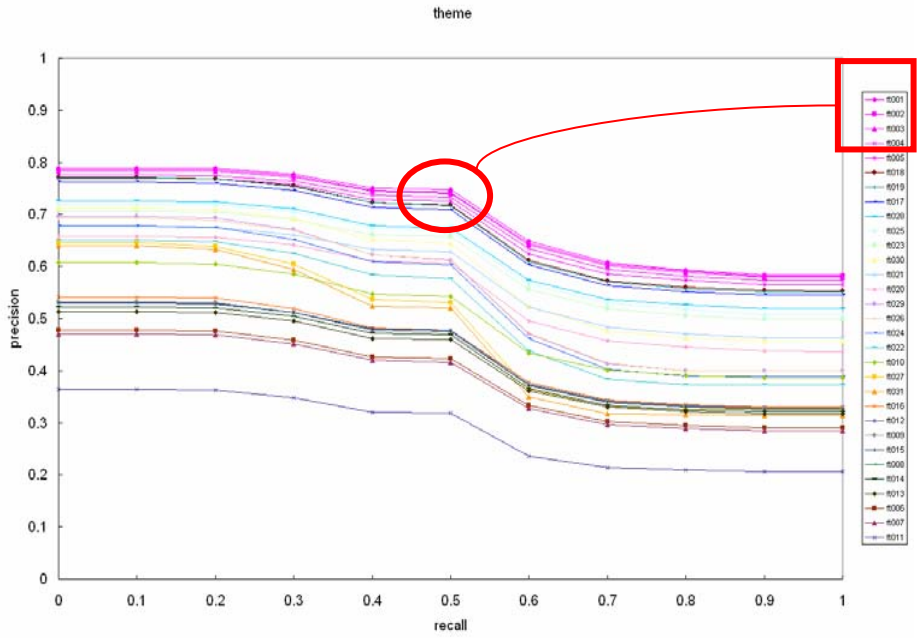
  - Threshold for selection
    - Fixed threshold = { t | $t \in [0, n]$ }
    - Floating threshold

      = { t | $t \in$[avg(*TFICF*(*w*(*d*)))-*m*, avg(*TFICF*(*w*(*d*)))+*m*]
      - *m, n* is an integer got from empirical experiments
- Classifier: MAXENT Toolkit

Document Categorization Subtask

- kNN-based Theme classification
  - Training documents: 2 years documents (1993, 1997)
  - Equally weighted merging
    - (technological field, technological field)
    - (purpose, purpose)
    - (method, method)

RunID	Condition	MAP
ft001	k=10	0.6872
ft002	k=20	0.6842
ft003	k=30	0.6819
ft004	k=50	0.6744
ft005	k=100	0.6666



Experiments on MEM-based Classification

- K-NN-based approach showed better MAP than MEM based one

Approach (ID)	Condition	Training data size (GB)	MAP	Top Map of k-NN
MEM (ft006)	TFICF, avg-2	1.12	0.3776	0.6872
MEM (ft007)	TFICF, avg	0.27	0.3709	

- MEM was more fit to F-Term classification than Theme one
  - MEM is relatively better for small category classification

Run	ID	Condition	MAP	Top MAP among all teams/runs
Theme	dt001	TFICF, avg-1	0.3776	0.6928 (dry run) 0.6872 (formal run)
	dt002	TFIDF, avg	0.3709	
F-term	df001	TFICF, avg-1	0.4819	0.4819 (dry run) 0.4998 (formal run)
	ffx01	TFICF, avg-2	0.4001	

Conclusions

- Component-by-component comparison
  - Re-organization by using structural information of patent
  - Precise comparison of components among document
- K-NN based approach for patent categorization
  - Elimination of un-important documents for noisy reduction
  - Using similar k documents for categorization