Patent Document Retrieval and Classification at KAIST

KAIST CS Dept. / BOLA

2005. 12. 7.

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

Contents

- Introduction
- Our approach for Retrieval and Categorization
  - Component-by-component retrieval
  - kNN-based approach
- Retrieval and Categorization System
- Experimental Results
- Conclusions

Introduction

- Patent Retrieval Task
  - Document retrieval subtask
  - **Theme categorization subtask**
    - Main topic in today's talk

- Our assumption
  - Only some parts of patent document are useful for retrieval and categorization

➔ To find important parts and to utilize them well

Characteristic of Patent Documents

❖ *Patent document is structured*

Normative section

Detailed component

Applicant-defined tags

	<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し 【請求項2】トラクタに対して3点リン
<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】

Usefulness of Detailed Components

- [prior art] and [application field]
  - Including much information related to technical background and technical field
    - They can be more helpful to categorize patent documents
- [purpose] and [means of solving problems]
  - Representing the whole patent document
  - Often used in the <Abstract> section
- <description> section
  - including many noises
    - ➔ Selecting useful detailed components in the section

Detailed components → Major features

Our Approach for Categorization

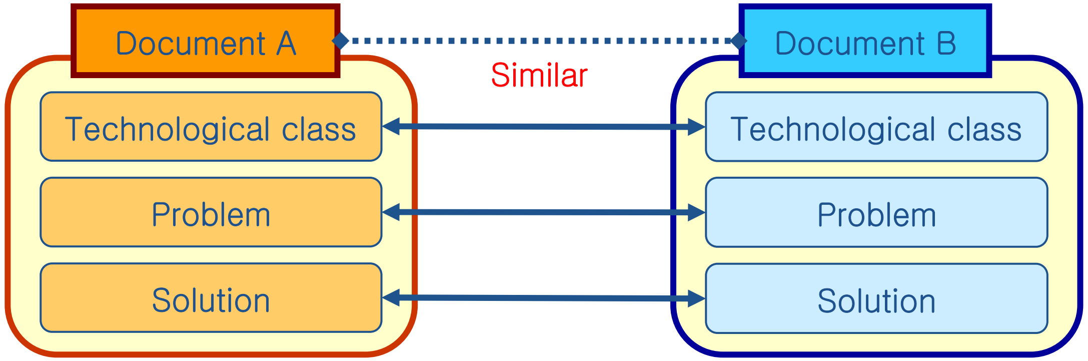
- kNN-based patent categorization
  - Retrieving k similar documents from training set
  - Classifying a given patent into the theme codes of *k* similar documents

- Motivation
  - Word-bag Vectors
    - are often used in many machine learning method
    - too large vocabulary size in a large-scaled training set

kNN-based approach

Our Approach for Retrieval

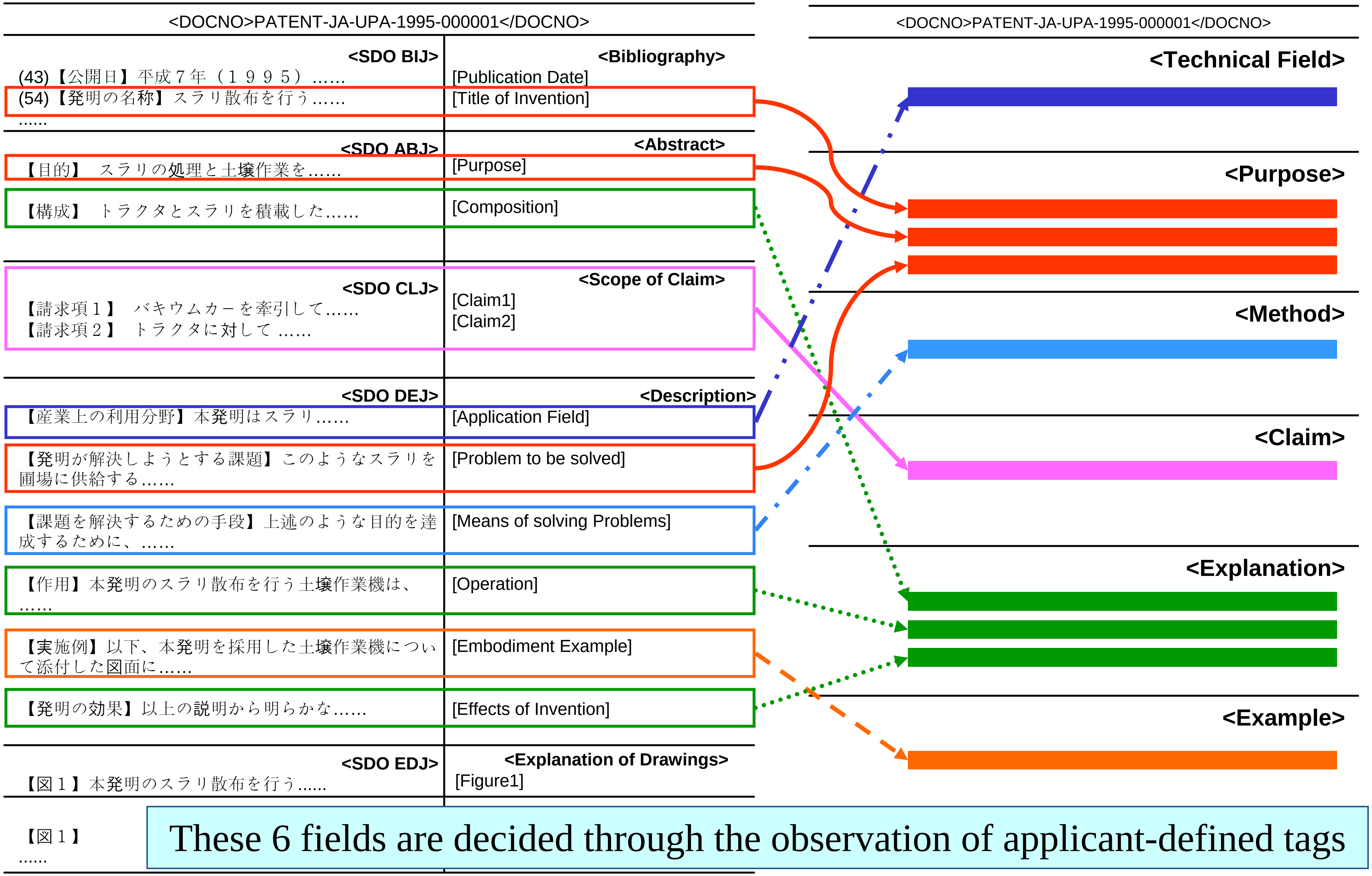
- Component-by-component comparison
  - If two documents are in the same technical classes and have the same problem and solution (method)
  - ➔ They are similar



- Motivation
  - Document–by-document comparison
    - Mixed information may cause error

Component-by-component comparison

Re-organization for component-by-component comparison



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

<SDO BIJ>

(43)【公開日】平成７年（１９９５）……

(54)【発明の名称】スラリ散布を行う……

……

<Bibliography>

[Publication Date]

[Title of Invention]

<SDO ABJ>

【目的】　スラリの処理と土壌作業を……

【構成】　トラクタとスラリを積載した……

<Abstract>

[Purpose]

[Composition]

<SDO CLJ>

【請求項１】　バキウムカーを牽引して……

【請求項２】　トラクタに対して ……

<Scope of Claim>

[Claim1]

[Claim2]

<SDO DEJ>

【産業上の利用分野】本発明はスラリ……

【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……

<Description>

[Application Field]

[Problem to be solved]

[Means of solving Problems]

[Operation]

[Embodiment Example]

[Effects of Invention]

<SDO EDJ>

【図１】本発明のスラリ散布を行う......

【図１】

……

<Explanation of Drawings>

[Figure1]

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

<Technical Field>

<Purpose>

<Method>

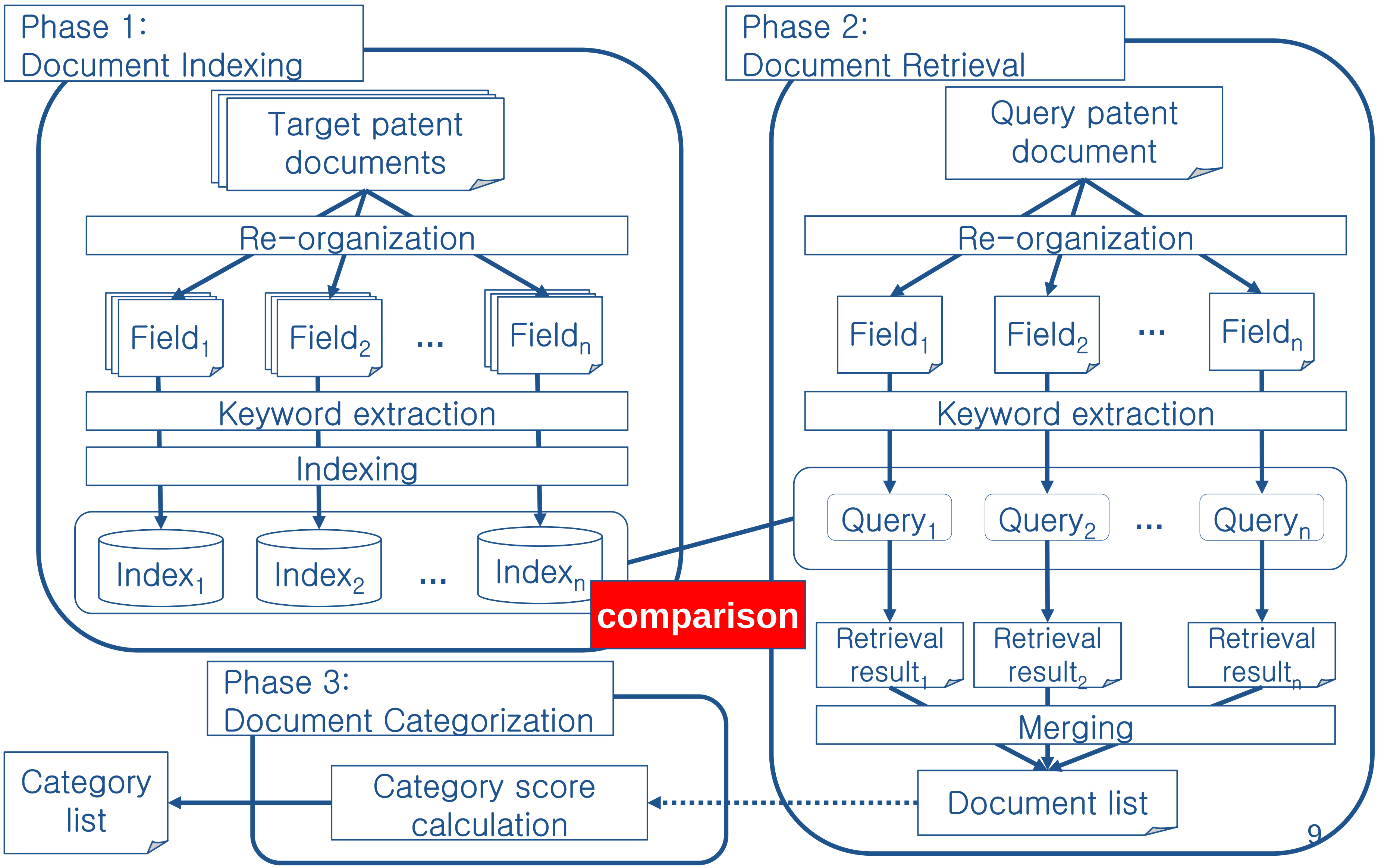
<Claim>

<Explanation>

<Example>

These 6 fields are decided through the observation of applicant-defined tags

System Architecture for Retrieval and Categorization



How to re-organize a document?

Phase 1: Indexing

Re-organizing by using applicant-defined tags

- Various tags in <ABJ>, <DEJ>
  - 3,516 tags (among 347,227 doc.)

Frequency	applicant defined tags (Japanese)	Meaning
310,276	課題を解決するための手段	Means of solving problems
2,502	問題点を解決するための手段	
1,449	課題を解決する為の手段	
2,923	課題を解決するための手段及び作用	Means of solving problems & Operation
3,962	従来の技術及び発明が解決しようとする課題	Prior art & Problem to be solved
306,350	発明が解決しようとする課題	Problem to be solved
2,121	発明が解決しようとする問題点	
1,476	発明が解決しようとしている課題	

- Classifying tags according to head nouns of tags
  - 100 most frequent HNs ➔ 6 classes

Examples of Classified Applicant-defined Tags

Phase 1: Indexing

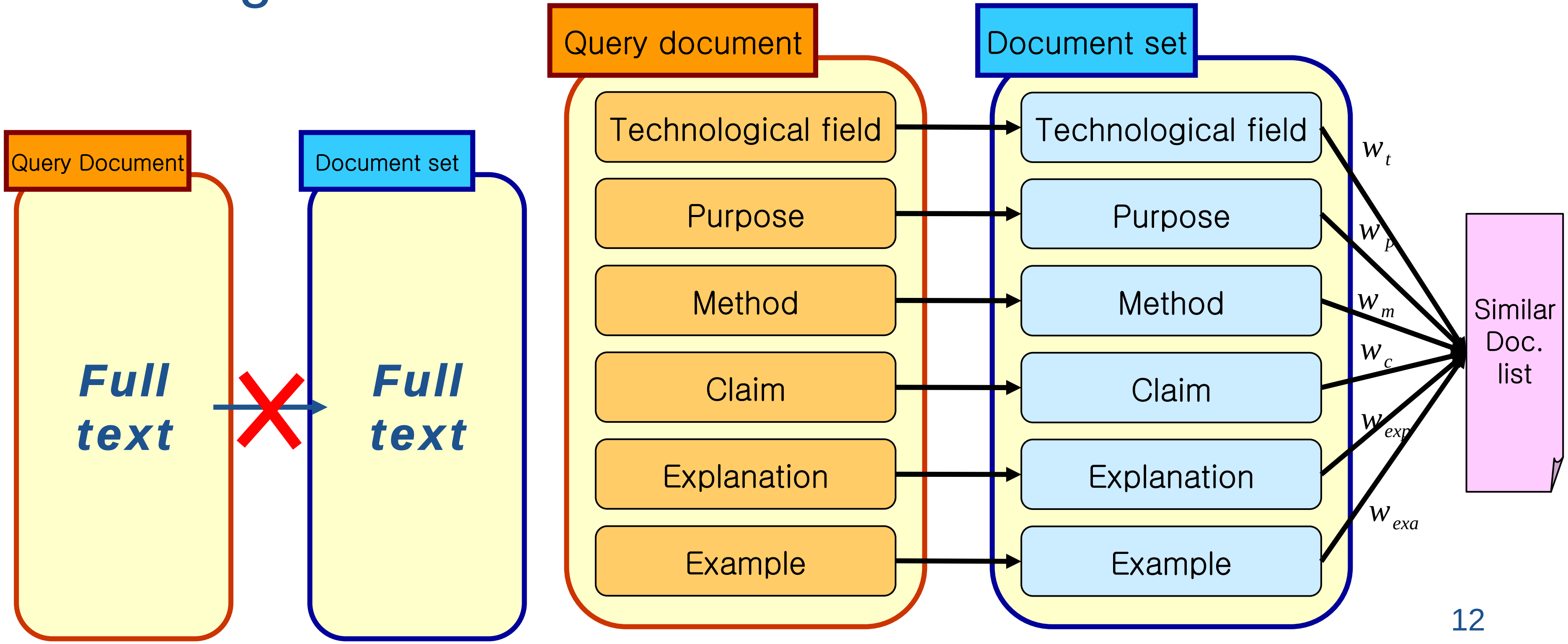
Primitive label	Head	Examples of Applicant –defined tags
Technological field	分野 (field) 技術 (art) 背景 (background)	産業上の利用分野 (Industrial application field) 従来の技術 (prior art) 発明の背景 (background of the invention)
Purpose	名称 (title) 目的 (purpose) 課題 (problem)	発明の名称 (title of the invention) 発明の目的 (purpose of the invention) 発明が解決しようとする課題 (problem to be solved by the invention)
Method	手段 (the means)	問題点を解決するための手段 (the means of solving the problem) 課題を解決するための手段及び作用 (the means of solving the problem and the operation)
Claim		All titles in the <Claim> part
Explanation	構成 (composition) 効果 (effect) 作用 (operation) 説明 (explanation)	構成 (Composition) 発明の効果 (the effect of the invention) 課題を解決するための手段及び作用 (the means of solving the problem and the operation) 発明の具体的説明 (The concrete explanation of composition)
Example	例 (example)	実施例 (embodiment example) 参考例 (referential example) 実験例 (experimental example)

Multiple classification

Document Retrieval

Phase 2: Retrieval

- Pairs of same fields are compared
  - 6 queries – 6 indexes
  - by Lemur toolkit
- Weighted summation



Assigning theme codes (1/2)

Phase 3: Categorization

Method

- by theme codes of *k* documents similar to a query document
  - Retrieved documents have theme codes
    - – Example) similarity result for a query document

rank	doc ID	document similarity	given theme codes
1	d04	371.773	2B062
2	d01	371.009	2B062, 2B304
3	d02	370.981	2B072
4	d03	370.863	2B304, 3L045, 3L055
5	d09	370.800	3L045
……			

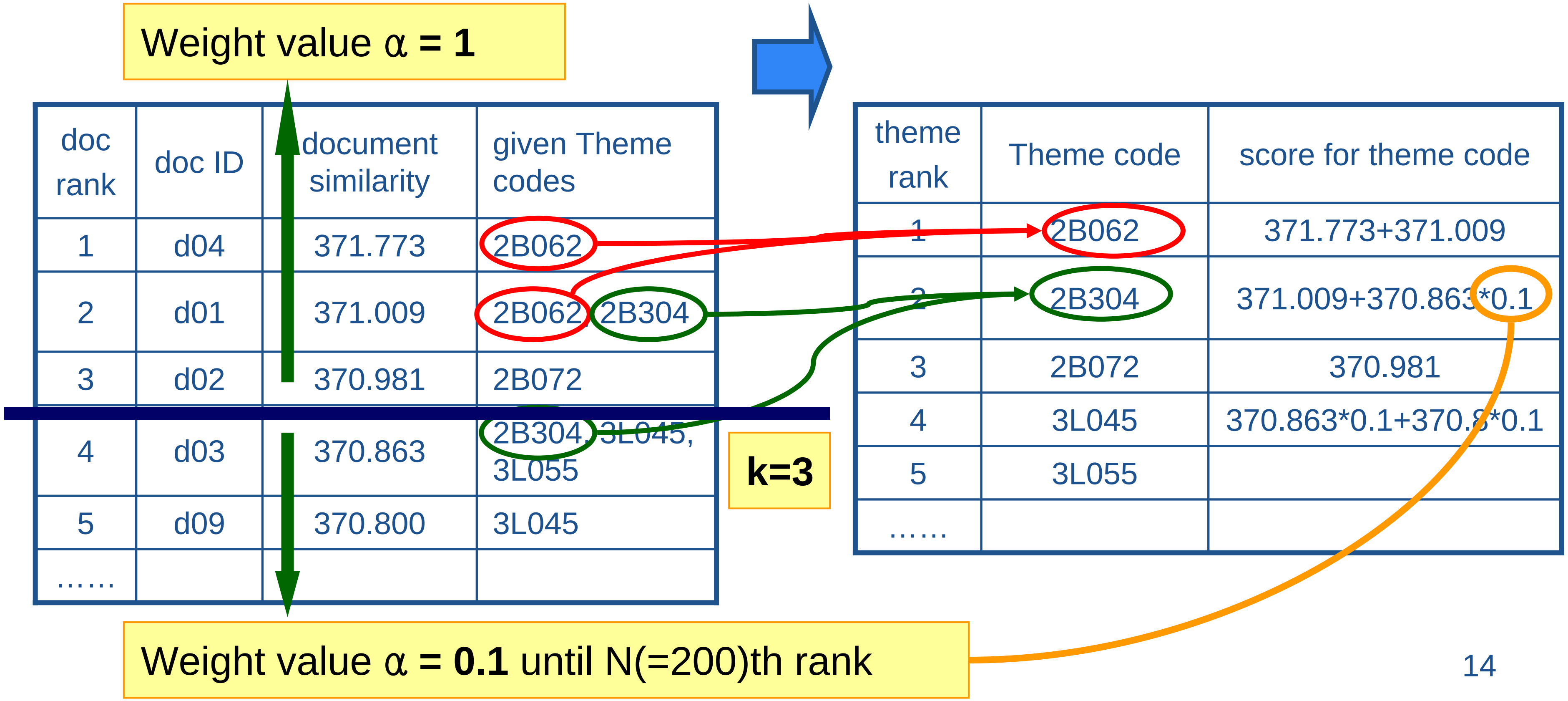
K=3 means that top 3 documents are meaningful among N retrieved documents

Assigning theme codes (2/2)

Phase 3: Categorization

- Method for calculating scores of theme codes

Example for a given query



Similarity Result

Weight value α = **1**

doc rank	doc ID	document similarity	given Theme codes
1	d04	371.773	2B062
2	d01	371.009	2B062, 2B304
3	d02	370.981	2B072
4	d03	370.863	2B304, 3L045, 3L055
5	d09	370.800	3L045
……			

k=3

Weight value α = **0.1** until N(=200)th rank

Score for Theme code

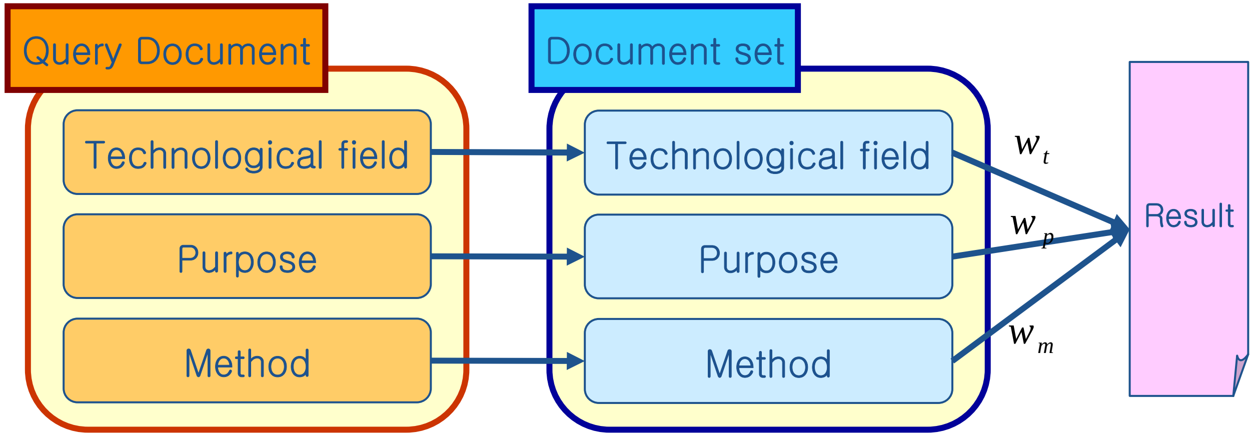
theme rank	Theme code	score for theme code
1	2B062	371.773+371.009
2	2B304	371.009+370.863*0.1
3	2B072	370.981
4	3L045	370.863*0.1+370.8*0.1
5	3L055	
……		

Experimental Results

Theme Categorization Subtask

❖ kNN-based Theme categorization

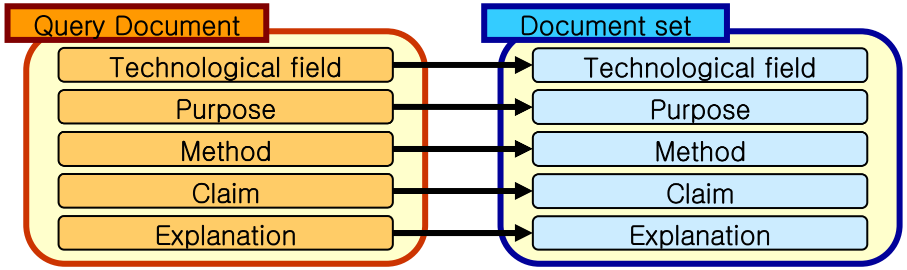
- Training documents: 2 years documents (1993, 1997)
- Component-by-component retrieval
- Equally weighted merging

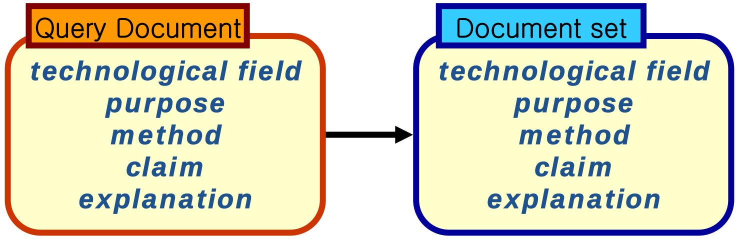


RunID	Condition	MAP	Rank
ft001	k=10	0.6872	1
ft002	k=20	0.6842	2
ft003	k=30	0.6819	3

Additional experiments

- kNN-based approach vs. Word-bag Vector
  - MAP in NTCIR-5 formal run
    - 0.6872 (kNN) > 0.3776 (MEM)
- Detailed component vs. Normative section
  - Dev. Set (1 year training data, 100 test documents)
    - 0.6372 (Purpose) > 0.5774 (Description section)
- Component-by-Component vs. Doc–by-Doc
  - Dev. Set
    - 0.6402 (Technical field, Purpose, Method, Claim, Explanation)
    - 0.6050 (Technical field+Purpose+Method+Claim+Explanation)





Conclusions

- kNN based approach for patent categorization
  - Using similar N documents
    - ➔ the effect of feature reduction (cf. word-bag vector)

- Component-by-component comparison
  - Considering meaningful small units in a document
    - ➔ precise comparison of content among documents
    - cf. document-by-document comparison