Document Structure Analysis in Associative Patent Retrieval

Atsushi Fujii, Tetsuya Ishikawa

University of Tsukuba

Invalidity search task

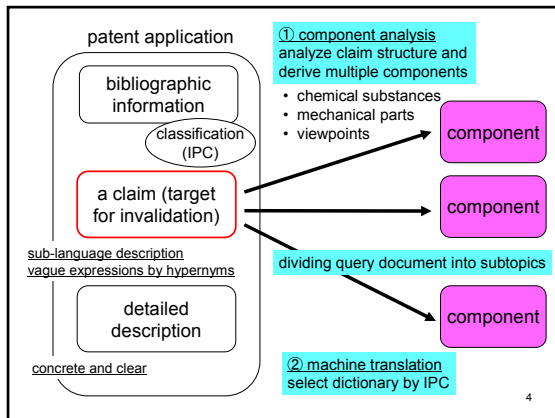- Find the patents that can invalidate the demand in a patent application (claim)
- This can be seen as patent-to-patent associative retrieval
  - both queries and documents are patents
- This task is usually performed by
  - examiners in a government patent office
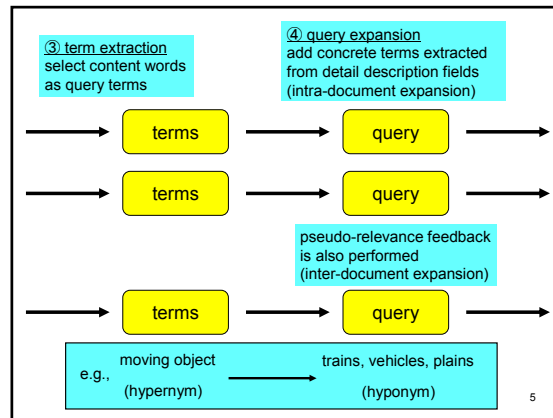 - searchers of IP division in private companies

2

Basis of our system

- because query is long, structure analysis is effective
1. analyze claim structure and divide query into subtopics (local structure analysis)
2. expand query terms using "detailed description" fields (global structure analysis)
3. search for documents on a subtopic-by-subtopic basis
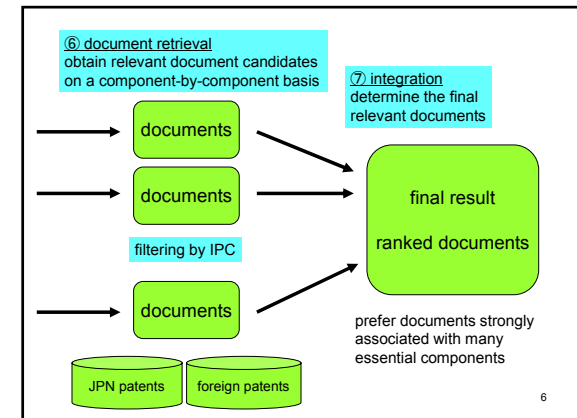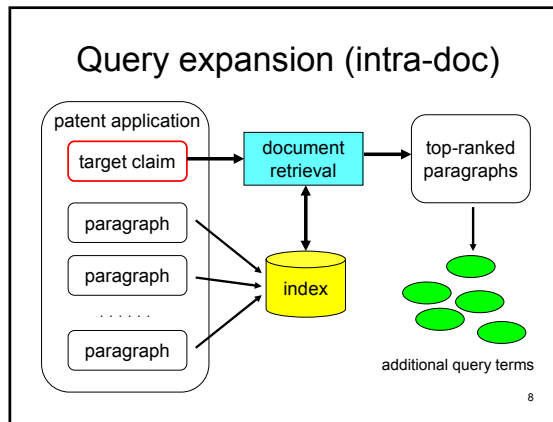4. integrate and re-rank documents

3

---

patent application

bibliographic information

classification (IPC)

a claim (target for invalidation)

sub-language description
vague expressions by hypernyms

detailed description

concrete and clear

① component analysis
analyze claim structure and derive multiple components
- chemical substances
- mechanical parts
- viewpoints

component

component

dividing query document into subtopics

component

② machine translation
select dictionary by IPC

4

---

③ term extraction
select content words as query terms

④ query expansion
add concrete terms extracted from detail description fields (intra-document expansion)

terms → query

terms → query

pseudo-relevance feedback is also performed (inter-document expansion)

terms → query

e.g., moving object → trains, vehicles, plains
(hypernym) → (hyponym)

5

---

⑥ document retrieval
obtain relevant document candidates on a component-by-component basis

⑦ integration
determine the final relevant documents

documents

documents

filtering by IPC

documents

final result

ranked documents

JPN patents   foreign patents

prefer documents strongly associated with many essential components

6

1

Example claim

components (subtopics)

	Claim #1	Candidate documents		
		A	B	C
1	移動体の周囲を監視するための移動体の周囲監視装置であって、	100	20	0
2	その周囲の領域の映像を光学像に中心射影変換する光学系と、撮像レンズを含み、前記中心射影変換された光学像を画像データに変換する撮像部とを含む少なくとも1つの全方位視覚センサーと、	40	10	100
3	前記画像データをパノラマ画像データおよび透視画像データの少なくとも一方に変換する画像処理部と、	10	0	30
⋮	⋮	⋮	⋮	⋮
7	前記表示部が、前記移動体の周囲を俯瞰する前記透視画像を表示する、移動体の周囲監視装置。	80	30	0
	weighted average of scores	41.4 (1)	17.1 (3)	27.1 (2)

relevance score

Query expansion (intra-doc)

patent application

target claim → document retrieval → top-ranked paragraphs

paragraph
paragraph
· · · · · ·
paragraph

→ index ↔ document retrieval

→ additional query terms

Example of intra-doc expansion

明細書
【請求項1】 移動体 の 周囲 を 監視 するための 移動体 の 周囲監視装置 であって、

【発明の詳細な説明】
【0001】
　・・・より詳細には、人 または 貨物 を 輸送 する 自動車 または 電車 を含む 移動体 の 周囲 を 監視 するために用いられる 移動体 の 周囲監視装置 に関する。

Comparative experiments

- Optional methods (yes/no)
  - A: component analysis
    - based on Japanese punctuation
  - B: intra-document expansion
  - C: character bigram index terms
    (combined with word index terms)
  - D: pseudo-relevance feedback
  - E: filtering by IPC

Results (MAP)

A	B	C	D	E	Rigid	Relax
yes	yes	yes	yes	yes	13.83	12.97
yes	yes	yes	yes		14.64	13.43
	yes	yes	yes		14.05	13.01
		yes	yes		14.60	13.23
yes	yes		yes		10.78	10.32
yes	yes	yes			14.46	12.96

Summary and future work

- the following methods were effective
  - claim component analysis
  - intra-document expansion
  - combination of word and character indexes
  - pseudo-relevance feedback
- the use of IPC was not effective
  - more sophisticated use is needed
- future work includes applying our method to other expository texts