A33　中国語キーワードによる中国特許情報解析

調査精度向上への応用

アジア特許情報研究会

〇花王株式会社　安藤俊幸
株式会社ＩＨＩ　金澤祐孝
電気化学工業株式会社　小山裕史
東ソー株式会社　沖 祥嘉

発表内容

中国語キーワードによる中国特許情報解析

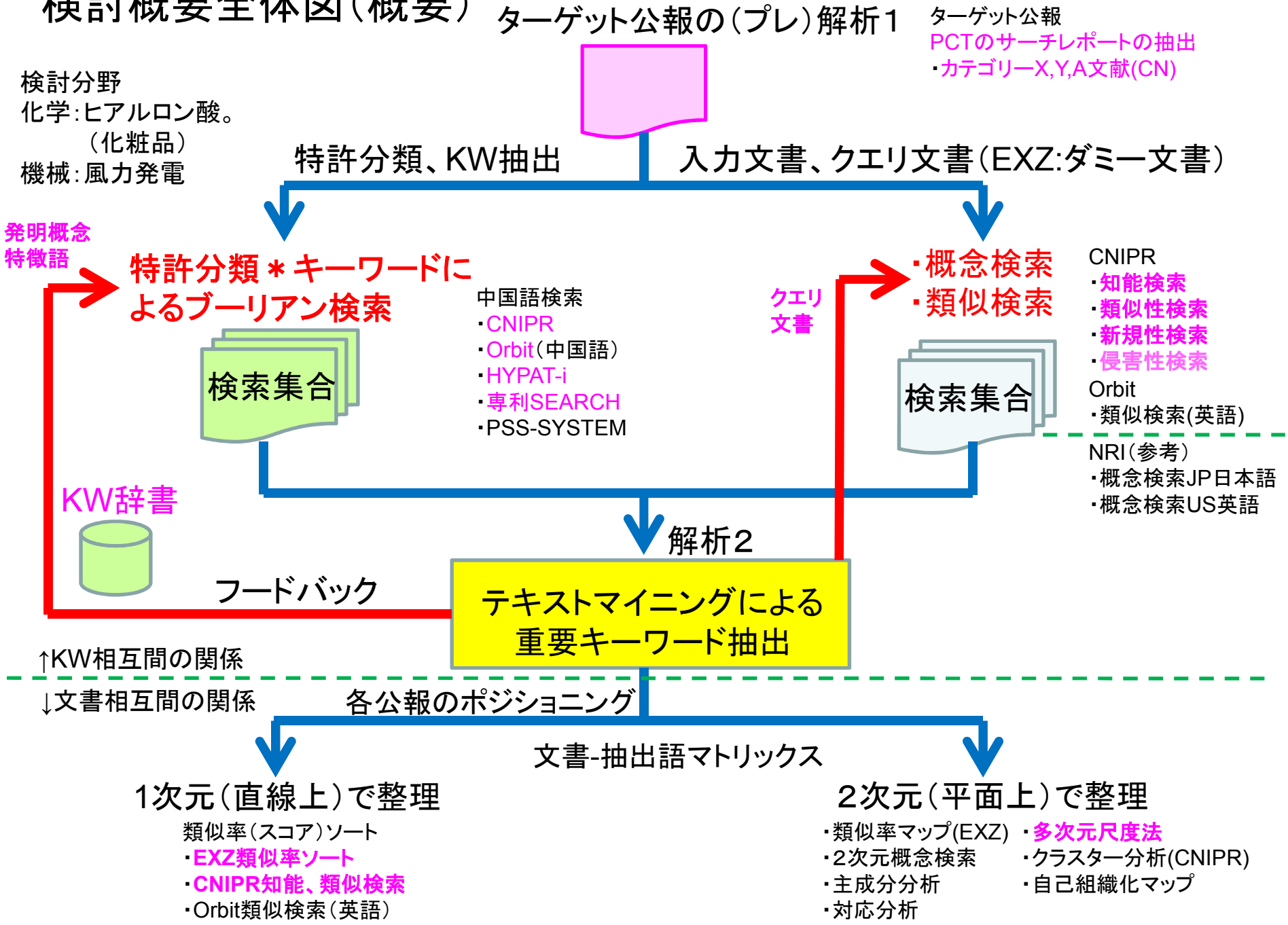
調査精度向上への応用

狙い　適合率向上による調査の効率化

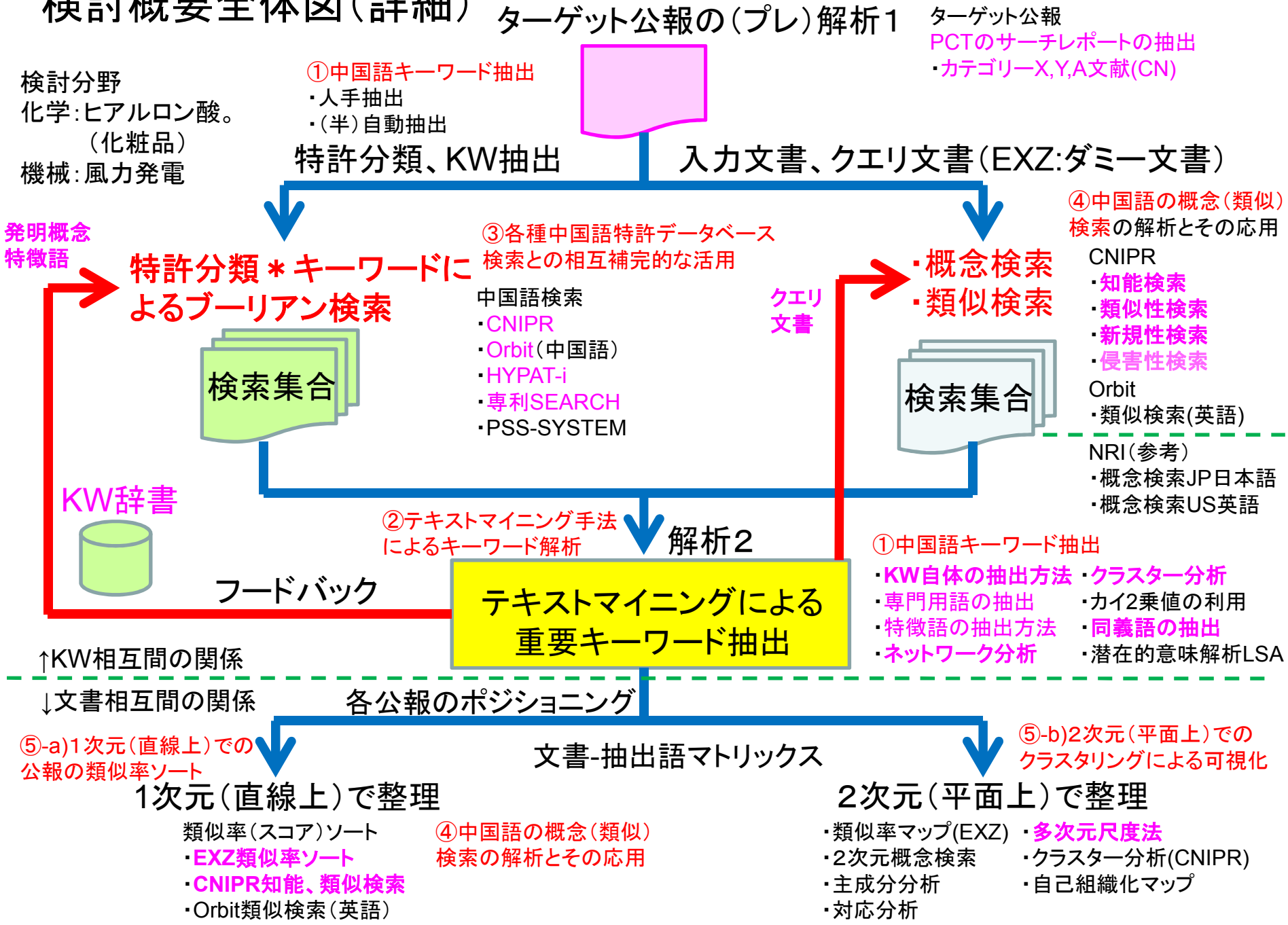
①複数の中国語キーワード抽出方法の比較検討
②テキストマイニング手法による重要キーワード、同義語抽出
③各種中国語特許データベース検索との相互補完的な活用
④中国語の概念（類似）検索の解析とその応用
⑤キーワード解析の応用として
・１次元（直線上）での公報の類似率ソート
・２次元（平面上）でのクラスタリングによる可視化

検討対象
化学分野：ヒアルロン酸の動向調査
機械分野：風力発電の出願前先行技術調査

検討概要全体図（概要）


ターゲット公報の（プレ）解析１
ターゲット公報
PCTのサーチレポートの抽出
・カテゴリーX,Y,A文献(CN)
検討分野
化学：ヒアルロン酸。
（化粧品）
機械：風力発電
特許分類、KW抽出
入力文書、クエリ文書（EXZ:ダミー文書）
発明概念
特徴語
特許分類＊キーワードによるブーリアン検索
検索集合
中国語検索
・CNIPR
・Orbit（中国語）
・HYPAT-i
・専利SEARCH
・PSS-SYSTEM
クエリ文書
・概念検索
・類似検索
検索集合
CNIPR
・知能検索
・類似性検索
・新規性検索
・侵害性検索
Orbit
・類似検索(英語)
NRI（参考）
・概念検索JP日本語
・概念検索US英語
KW辞書
解析２
フードバック
テキストマイニングによる重要キーワード抽出
↑KW相互間の関係
↓文書相互間の関係
各公報のポジショニング
文書-抽出語マトリックス
1次元（直線上）で整理
類似率（スコア）ソート
・EXZ類似率ソート
・CNIPR知能、類似検索
・Orbit類似検索（英語）
2次元（平面上）で整理
・類似率マップ(EXZ)
・2次元概念検索
・主成分分析
・対応分析
・多次元尺度法
・クラスター分析(CNIPR)
・自己組織化マップ

検討概要全体図（詳細）


ターゲット公報の（プレ）解析１
ターゲット公報
PCTのサーチレポートの抽出
・カテゴリーX,Y,A文献(CN)
検討分野
化学：ヒアルロン酸。
（化粧品）
機械：風力発電
①中国語キーワード抽出
・人手抽出
・（半）自動抽出
特許分類、KW抽出
入力文書、クエリ文書（EXZ:ダミー文書）
④中国語の概念（類似）
検索の解析とその応用
発明概念
特徴語
特許分類＊キーワードによるブーリアン検索
③各種中国語特許データベース
検索との相互補完的な活用
中国語検索
・CNIPR
・Orbit（中国語）
・HYPAT-i
・専利SEARCH
・PSS-SYSTEM
クエリ
文書
・概念検索
・類似検索
CNIPR
・知能検索
・類似性検索
・新規性検索
・侵害性検索
Orbit
・類似検索(英語)
検索集合
検索集合
NRI（参考）
・概念検索JP日本語
・概念検索US英語
KW辞書
②テキストマイニング手法
によるキーワード解析
解析２
①中国語キーワード抽出
・KW自体の抽出方法
・専門用語の抽出
・特徴語の抽出方法
・ネットワーク分析
・クラスター分析
・カイ2乗値の利用
・同義語の抽出
・潜在的意味解析LSA
フードバック
テキストマイニングによる
重要キーワード抽出
↑KW相互間の関係
↓文書相互間の関係
各公報のポジショニング
⑤-a)1次元（直線上）での
公報の類似率ソート
文書-抽出語マトリックス
⑤-b)2次元（平面上）での
クラスタリングによる可視化
1次元（直線上）で整理
類似率（スコア）ソート
・EXZ類似率ソート
・CNIPR知能、類似検索
・Orbit類似検索（英語）
④中国語の概念（類似）
検索の解析とその応用
2次元（平面上）で整理
・類似率マップ(EXZ)
・2次元概念検索
・主成分分析
・対応分析
・多次元尺度法
・クラスター分析(CNIPR)
・自己組織化マップ

①中国語キーワード抽出方法

抽出方法	説明	特徴	Web利用
ICTCLAS [1)]	中国語形態素解析ツール	品詞情報出力	○
IKAnalyzerNet [2)]	C#から呼び出し可能な 中国語分詞ライブラリ	KWの位置情報出力 部分的N-gram出力	
言選Web（中文版）[3)]	専門用語（キーワード） 自動抽出サービス	専門用語抽出、 多言語対応	○
パテントマップEXZ	パテントマップソフトの 組み込み機能	日本語、英語、中国語 キーワード切り出し可	
Orbit.com	中国語KW分離制御コード （不可視）を利用した抽出	中国語分詞ソフトにより 分離していると思われる	
Microsoft Word 中国語版	Wordの組み込み機能	VBAマクロより利用可能	

1)ICTCLAS：**I**nstitute of **C**omputing **T**echnology,**C**hinese **L**exical **A**nalysis **S**ystem
階層的隠れマルコフモデルを使用した中国語形態素解析ツール
http://ictclas.nlpir.org/

2)IKAnalyzerNet
http://www.piaoyi.org/c-sharp/IKAnalyzerNet.html

3)言選Web（中文版）
http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_cn.html

形態素解析とは

（参考）和布蕪による区切り　　文章を、形態素（言語で意味を持つ最小単位）に分割すること

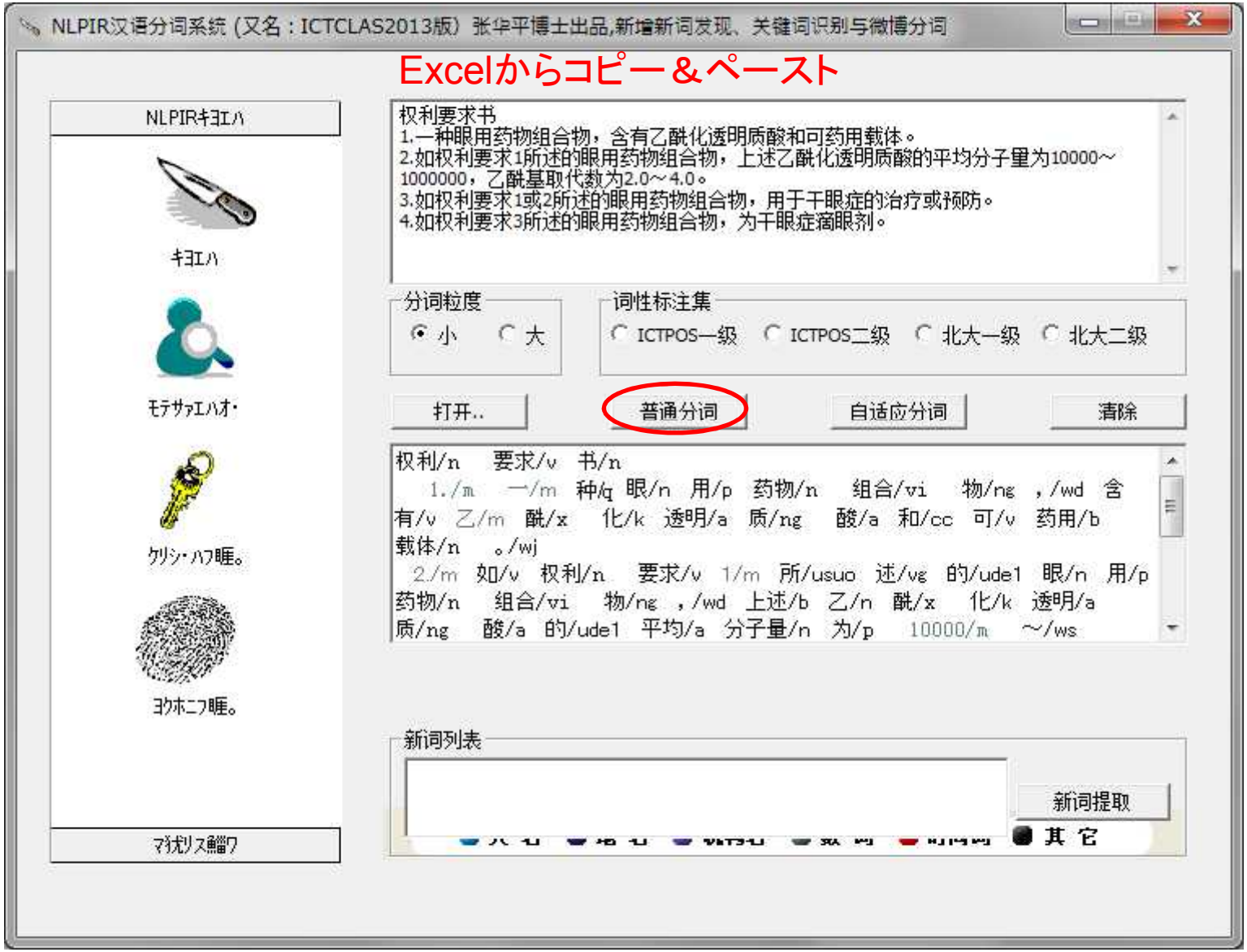
1. アセチル化ヒアルロン酸と医薬的に許容される担体とを含有する、眼用医薬組成物。

JP2004262777A→CN1753913A対応特許

```
1	名詞,数,*,*,*,*,*
.	名詞,サ変接続,*,*,*,*,*
アセチル	名詞,一般,*,*,*,*,*
化	名詞,接尾,サ変接続,*,*,*,化,カ,カ
ヒアルロン	名詞,一般,*,*,*,*,*
酸	名詞,一般,*,*,*,*,酸,サン,サン
と	助詞,並立助詞,*,*,*,*,と,ト,ト
医薬	名詞,一般,*,*,*,*,医薬,イヤク,イヤク
的	名詞,接尾,形容動詞語幹,*,*,*,的,テキ,テキ
に	助詞,副詞化,*,*,*,*,に,ニ,ニ
許容	名詞,サ変接続,*,*,*,*,許容,キョヨウ,キョヨー
さ	動詞,自立,*,*,サ変・スル,未然レル接続,する,サ,サ
れる	動詞,接尾,*,*,一段,基本形,れる,レル,レル
担体	名詞,一般,*,*,*,*,*
と	助詞,並立助詞,*,*,*,*,と,ト,ト
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
含有	名詞,サ変接続,*,*,*,*,含有,ガンユウ,ガンユー
する	動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
、	記号,読点,*,*,*,*,、,、,、
眼	名詞,一般,*,*,*,*,眼,メ,メ
用	名詞,接尾,一般,*,*,*,用,ヨウ,ヨー
医薬	名詞,一般,*,*,*,*,医薬,イヤク,イヤク
組成	名詞,サ変接続,*,*,*,*,組成,ソセイ,ソセイ
物	名詞,接尾,一般,*,*,*,物,ブツ,ブツ
。	記号,句点,*,*,*,*,。,。,。
EOS	（End Of Sentence）
```

MeCab (和布蕪)
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

ICTCLAS2013（NLPIR）による分詞



注）Windowsのシステムロケールを中国語に設定。

ICTCLAS2013（NLPIR）による分詞

CN1753913A

权利要求书　　**アセチル化ヒアルロン酸**と医薬的に許容される担体とを含有する、眼用医薬組成物。

1.一种眼用药物组合物，含有**乙酰化透明质酸**和可药用载体。

2.如权利要求1所述的眼用药物组合物，上述乙酰化透明质酸的平均分子量为10000～1000000，乙酰基取代数为2.0～4.0。

3.如权利要求1或2所述的眼用药物组合物，用于干眼症的治疗或预防。

4.如权利要求3所述的眼用药物组合物，为干眼症滴眼剂。

代表的な品詞

n:名词（名詞）	n:名词（名詞）	m:数词
v:动词（動詞）	vi:不及物动词（自動詞）	x:字符串（文字列）
m:数词（数詞）	cc:并列连词（調整接続詞）	k:后缀（接尾語）
q:量词（数量詞）	b:区别词（区別詞）	a:形容词
p:介词（前置詞）	u:助词（助詞）	ng:名词性语素（名詞形態素）

普通分詞

权利/n 要求/v 书/n

1./m 一/m 种/q 眼/n 用/p 药物/n 组合/vi 物/ng ，/wd 含有/v **乙/m 酰/x 化/k 透明/a 质/ng 酸/a** 和/cc 可/v 药用/b 载体/n 。/wj

2./m 如/v 权利/n 要求/v 1/m 所/usuo 述/vg 的/ude1 眼/n 用/p 药物/n 组合/vi 物/ng ，/wd 上述/b 乙/n 酰/x 化/k 透明/a 质/ng 酸/a 的/ude1 平均/a 分子量/n 为/p 10000/m ～/ws 1000000/m ，/wd 乙/m 酰/x 基/ng 取代/v 数/n 为/p 2.0/m ～/ws 4.0/m 。/wj

3./m 如/v 权利/n 要求/v 1/m 或/c 2/m 所/usuo 述/vg 的/ude1 眼/n 用/p 药物/n 组合/vi 物/ng ，/wd 用于/v 干眼症/n 的/ude1 治疗/vn 或/c 预防/v 。/wj

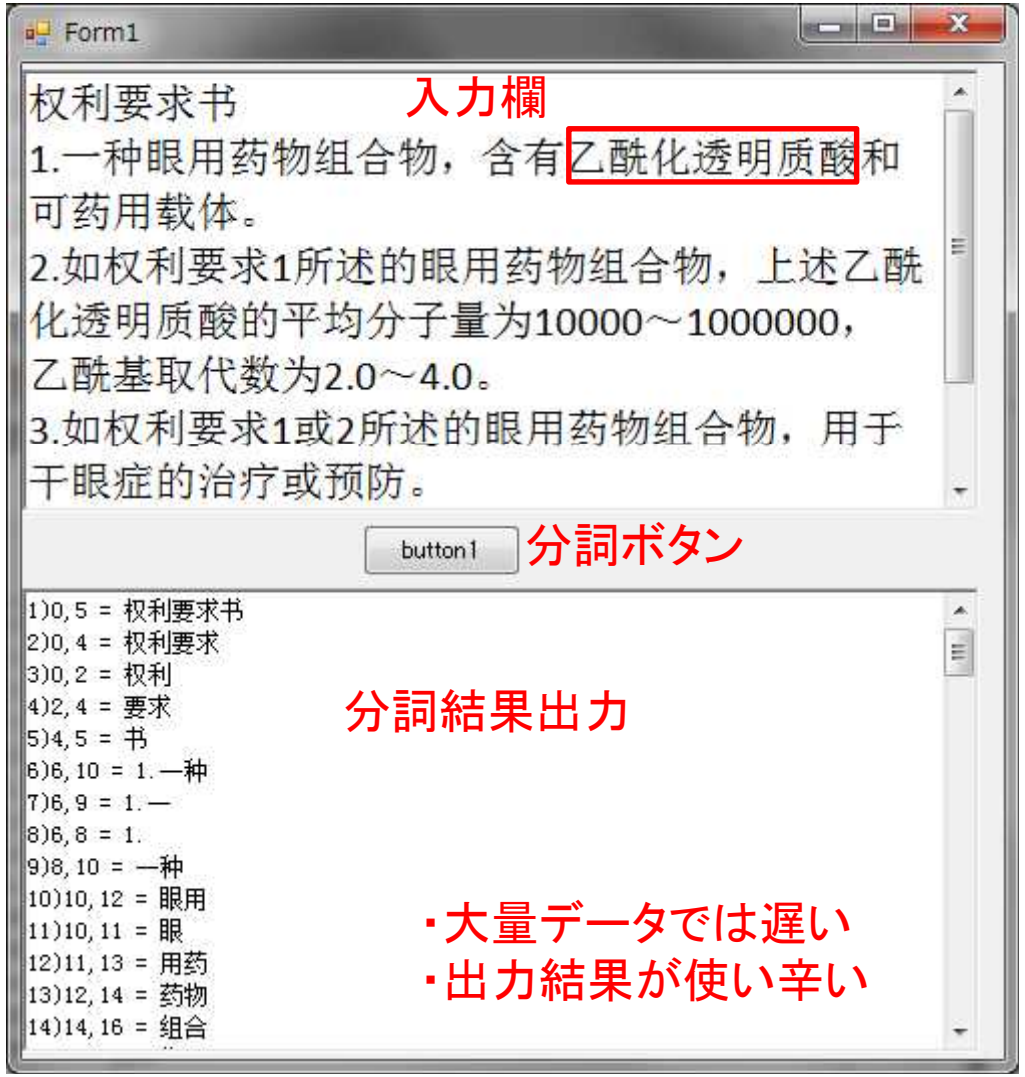
4./m 如/v 权利/n 要求/v 3/m 所/usuo 述/vg 的/ude1 眼/n 用/p 药物/n 组合/vi 物/ng ，/wd 为/p 干眼症/n 滴/q 眼/n 剂/q 。/wj

IKAnalyzerNetによる分詞

CN1753913A　**アセチル化ヒアルロン酸**と医薬的に許容される担体とを含有する、眼用医薬組成物。

权利要求书

1.一种眼用药物组合物，含有**乙酰化透明质酸**和可药用载体。



入力欄

分詞ボタン

分詞結果出力

・大量データでは遅い
・出力結果が使い辛い

IKAnalyzerNetに添付の**サンプルアプリ**「my3」

1)0,5 = 权利要求书
2)0,4 = 权利要求
3)0,2 = 权利
4)2,4 = 要求
5)4,5 = 书
6)6,10 = 1.一种
7)6,9 = 1.一
8)6,8 = 1.
9)8,10 = 一种
10)10,12 = 眼用
11)10,11 = 眼
12)11,13 = 用药
13)12,14 = 药物
14)14,16 = 组合
15)16,17 = 物
16)18,20 = 含有
17)20,23 = **乙酰化**
18)20,22 = **乙酰**
19)21,23 = **酰化**
20)23,26 = **透明**质
21)23,25 = **透明**
22)26,27 = **酸**
23)29,31 = 药用
24)31,33 = 载体
25)32,33 = 体

位置情報出力（文字数）

部分的N-gram

「言选Web」(中文版)による専門用語抽出

http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_cn.html

CN1753913
权利要求书
1.一种眼用药物组合物，含有乙酰化透明质酸和可药用载体。
2.如权利要求1所述的眼用药物组合物，上述乙酰化透明质酸的平均分子量为10000～1000000，乙酰基取代数为2.0～4.0。
3.如权利要求1或2所述的眼用药物组合物，用于干眼症的治疗或预防。
4.如权利要求3所述的眼用药物组合物，为干眼症滴眼剂。

Stop-list
药物组合物
如权利要求
所述的眼
权利要求书
乙酰化透明质酸和
干眼症滴眼剂
述乙酰化透明质酸的平均分子量
一种眼
干眼症的治疗
乙酰基取代数
预防
载体

ICTCLAS
权利
药物
干眼症
平均分子量
干眼症的治疗
眼剂
药用载体

パテントマップEXZによるキーワード抽出

CN1753913
权利要求书
1.一种眼用药物组合物，含有乙酰化透明质酸和可药用载体。
2.如权利要求1所述的眼用药物组合物，上述乙酰化透明质酸的平均分子量为10000～1000000，乙酰基取代数为2.0～4.0。
3.如权利要求1或2所述的眼用药物组合物，用于干眼症的治疗或预防。
4.如权利要求3所述的眼用药物组合物，为干眼症滴眼剂。

抽出キーワード

10000～1000000	数値範囲
2.0～4.0	
乙酰化透明质酸	アセチル化ヒアルロン酸
乙酰基取代数	アセチル基置換数
可药用载体	医薬的に許容される担体
干眼症	ドライアイ
干眼症滴眼剂	ドライアイ点眼剤
眼用药物组合物	眼用医薬組成物
治疗	治療又
平均分子量	平均分子量
权利	権利
预防	予防

抽出されなくても良いのでは？

Orbit.comの中国語、日本語の区切り記号

区切り記号（16進表記&H200b）不可視　→　□に置換して可視化

CN1753913Aダウンロードデータ

(CN1753913)
1. 一种眼用药物组合物，含有乙酰化透明质酸和可药用载体。

(CN1753913)
1. 一种□眼用□药物□组合□物，含有□乙酰□化□透明质酸□和□可□药用□载体。

JP2004262777A表示データ

1. アセチル化ヒアルロン酸と医薬的に許容される担体とを含有する、眼用医薬組成物。

1.アセチル□化□ヒアルロン□酸□と□医薬□的□に□許容□さ□れる□担体□
と□を□含有□する、眼□用□医薬□組成□物。

（参考）和布蕪による区切り

1/./アセチル/化/ヒアルロン/酸/と/医薬/的/に/許容/さ/れる/担体/
と/を/含有/する/、/眼/用/医薬/組成/物/。

中国語キーワード抽出結果まとめ

JP2004262777A

1. アセチル化ヒアルロン酸と医薬的に許容される担体とを含有する、眼用医薬組成物。

CN1753913Aの請求項1	1.一· 眼用药物组合物，含有乙酰化透明质酸和可药用载体。	赤字KWの分解数	特徴
ICTCLAS	1./m 一/m · /q 眼/n 用/p 药物/n 组合/vi 物/ng ，/wd 含有/v 乙/m 酰/x 化/k 透明/a 质/ng 酸/a 和/cc 可/v 药用/b 载体/n 。/wj	6	品詞情報
IKAnalyzer Net 一部抜粋	12)14,17 = 乙酰化 13)14,16 = 乙酰 14)15,17 = 酰化 15)17,20 = 透明质 16)17,19 = 透明 17)20,21 = 酸	6 N-gram除く 3	部分的N-gram 位置情報
パテントマップEXZ	乙酰化透明质酸 眼用药物组合物	1	専門用語
Orbit.com 注1）	1. 一· □眼用□药物□组合□物，含有□乙酰□化□透明质酸□和□可□药用□载体。	3	不可視

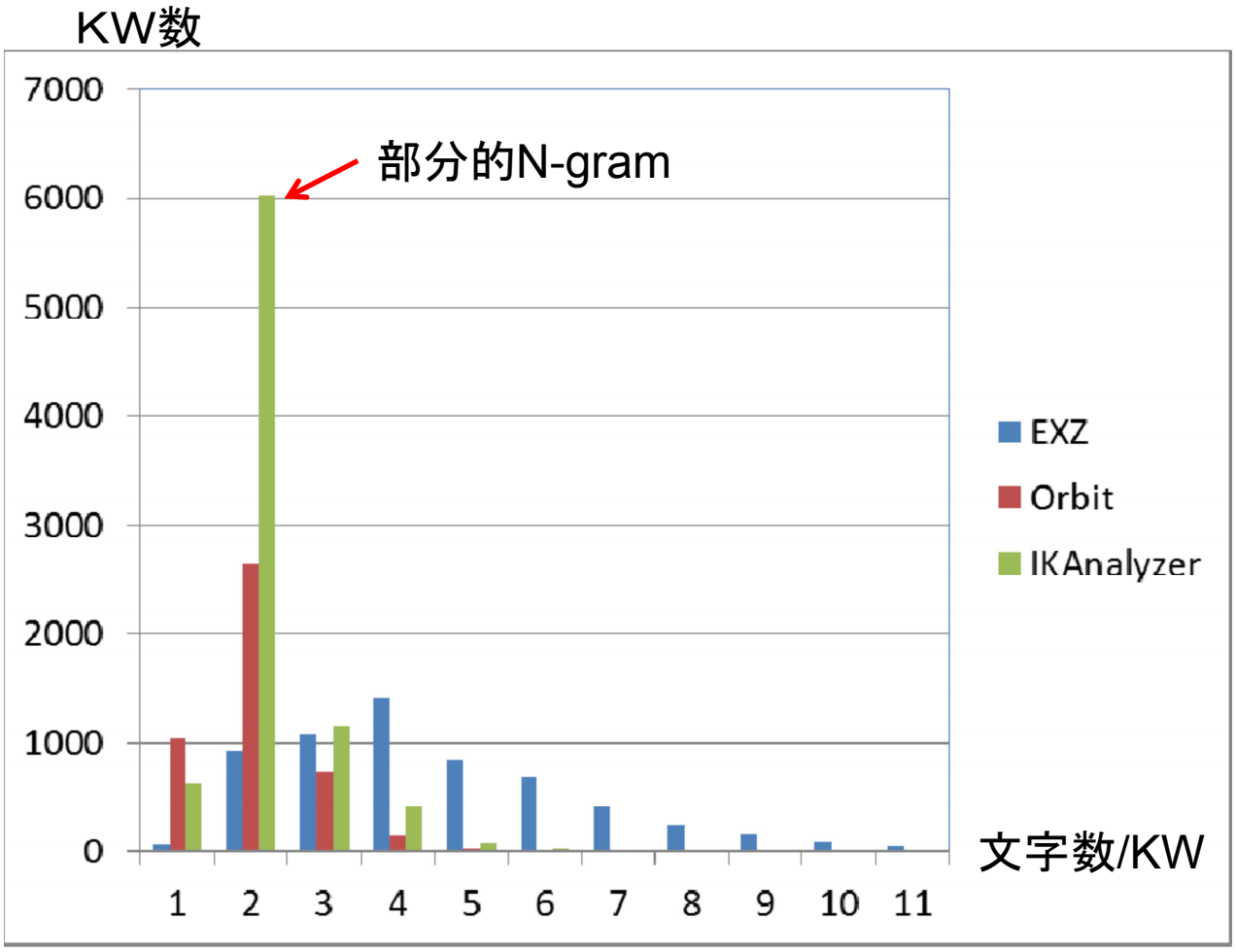
注1）不可視の区切り記号&H200bを□に置換して可視化

適合率向上要因：専門用語、分詞結果文字数→多い
網羅性向上要因：KW分解数→多い、部分的N-gram

品詞情報

中国語KW抽出方法別KWの文字数分布

IPC=A61K31/728(ヒアルロン酸)の検索結果200件（請求項）より抽出したKWの文字数分布



KW 文字数	EXZ KW	Orbit KW	IKAnalyzer KW	IKAnalyzer 頻度
1	74	1046	631	21473
2	923	2645	6031	73158
3	1074	735	1158	10276
4	1406	153	410	4934
5	856	26	90	608
6	687	11	22	132
7	410	3	6	27
8	244		0	0
9	154		1	5
10	96		1	1
11	51		2	5
途中略				
18	3			
計	6059	4619	8352	110619

CNIPRの概念検索による同義語、関連概念

http://search.cnipr.com/pages!advSearch.action

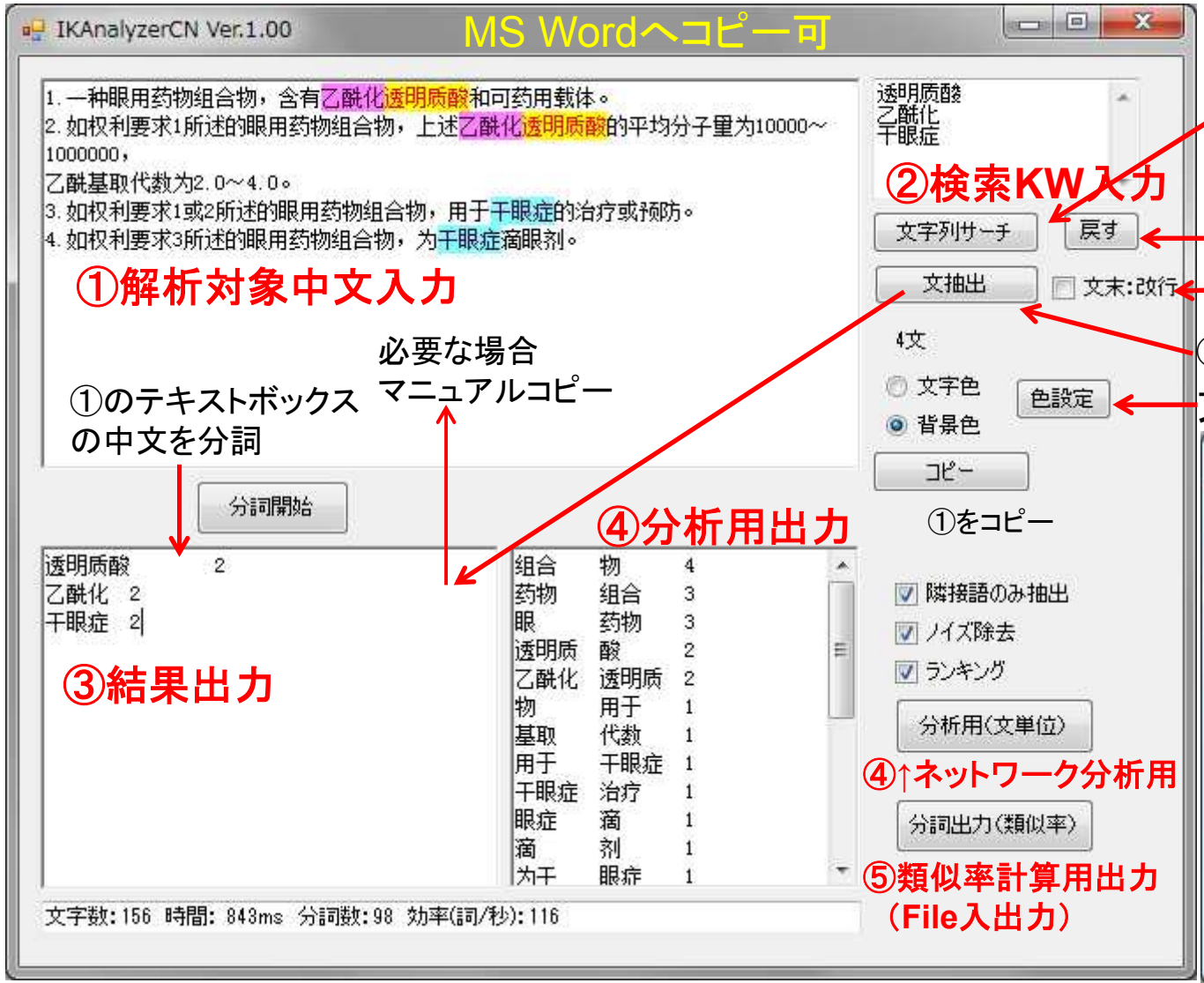
ヒアルロン酸の同義語、類義語の抽出結果

No.	日本語	中国語	CNIPR	Orbit CN指定	HYPAT-i
1	ヒアルロン酸	质酸	4100	4327	4114
2	ヒアルロン酸	玻尿酸	59	151	55
3	ヒアルロン酸	透明質酸	0	43	0
4	ヒアルロン酸	透明质酸	3688	3667	3640
5	ヒアルロン酸	玻璃酸	228	257	230
6	ウロン酸	糖醛酸	823	1717	1723
7	ヒアルロン酸Na	质酸钠	650	646	632
8	ヒアルロン酸Na	玻璃酸钠	183	186	182
9	ヒアルロン酸Na	透明质酸钠	617	598	607

対象：CN公開、TI+AB+CLM　　検索：2013.03.06

IKAnalyzerNetを改良/機能追加（IKAnalyzerCN）

C#言語でプログラミング



①のKWを文字列サーチして
カラー設定

①の文字色と背景色を元に戻す

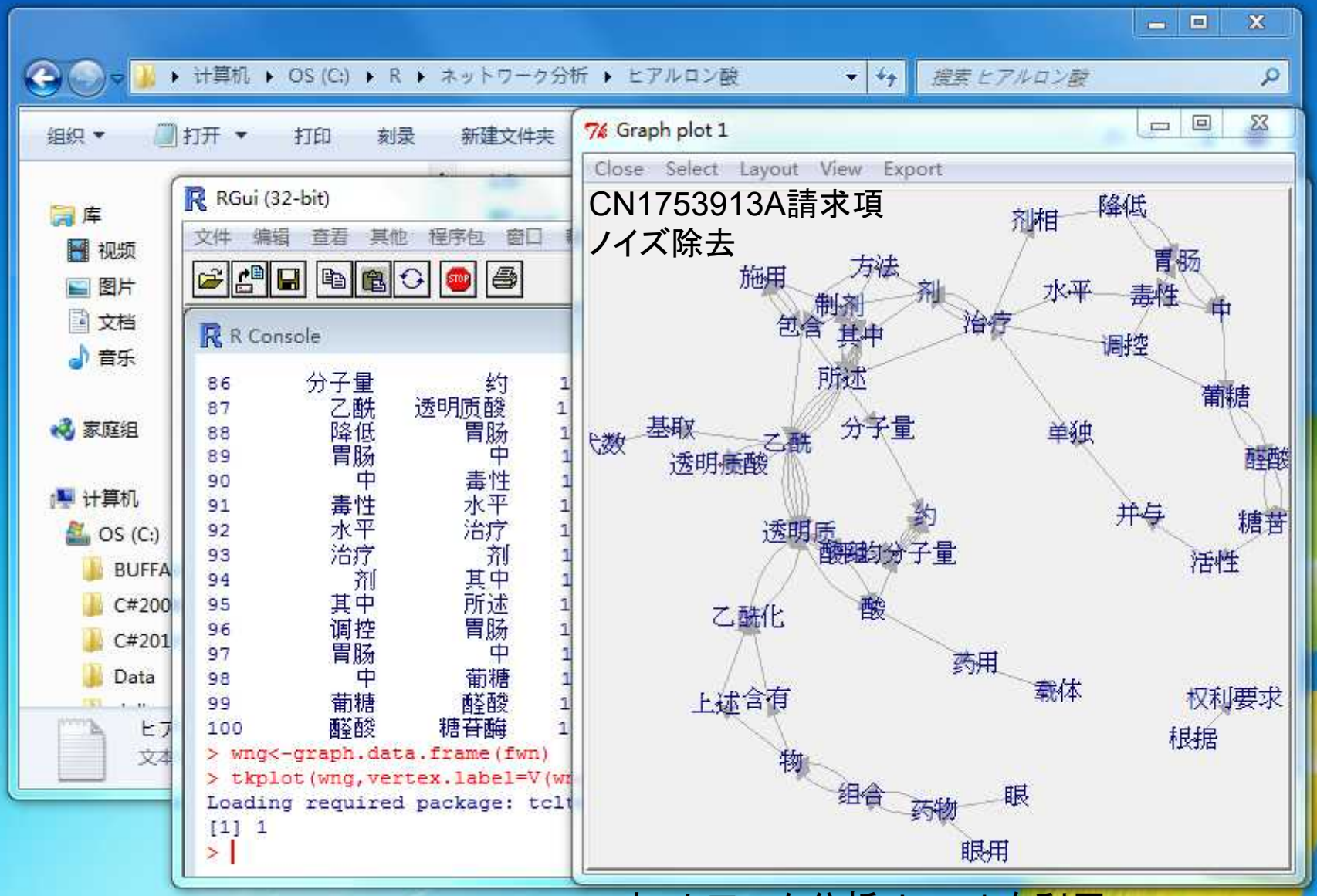
①の改行を文末とみなして抽出

②のKWを含む文を①→③に抽出

文字色と背景色の色設定

文字列サーチ機能とKW抽出機能のハイブリッド活用

中国語Windows+Rによるネットワーク分析



Windows 7 Ultimate：35言語に対応
コントロールパネルで、あるいは、ユーザーID毎に言語を変更可能

ネットワーク分析：igraphを利用

請求項のネットワーク分析

CN1753913A

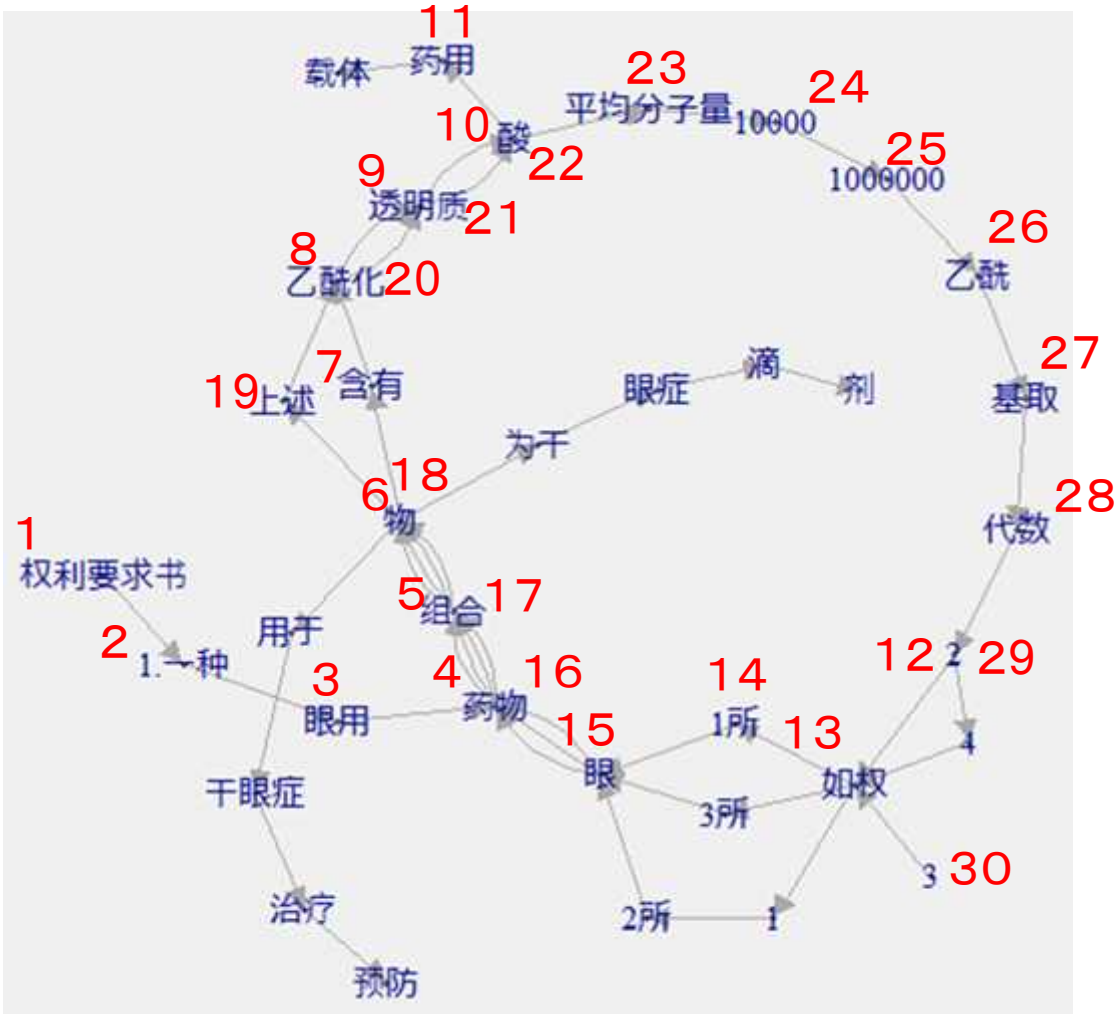
权利要求书

1.一种眼用药物组合物，含有乙酰化透明质酸和可药用载体。

2.如权利要求1所述的眼用药物组合物，上述乙酰化透明质酸的平均分子量为10000～1000000，乙酰基取代数为2.0～4.0。

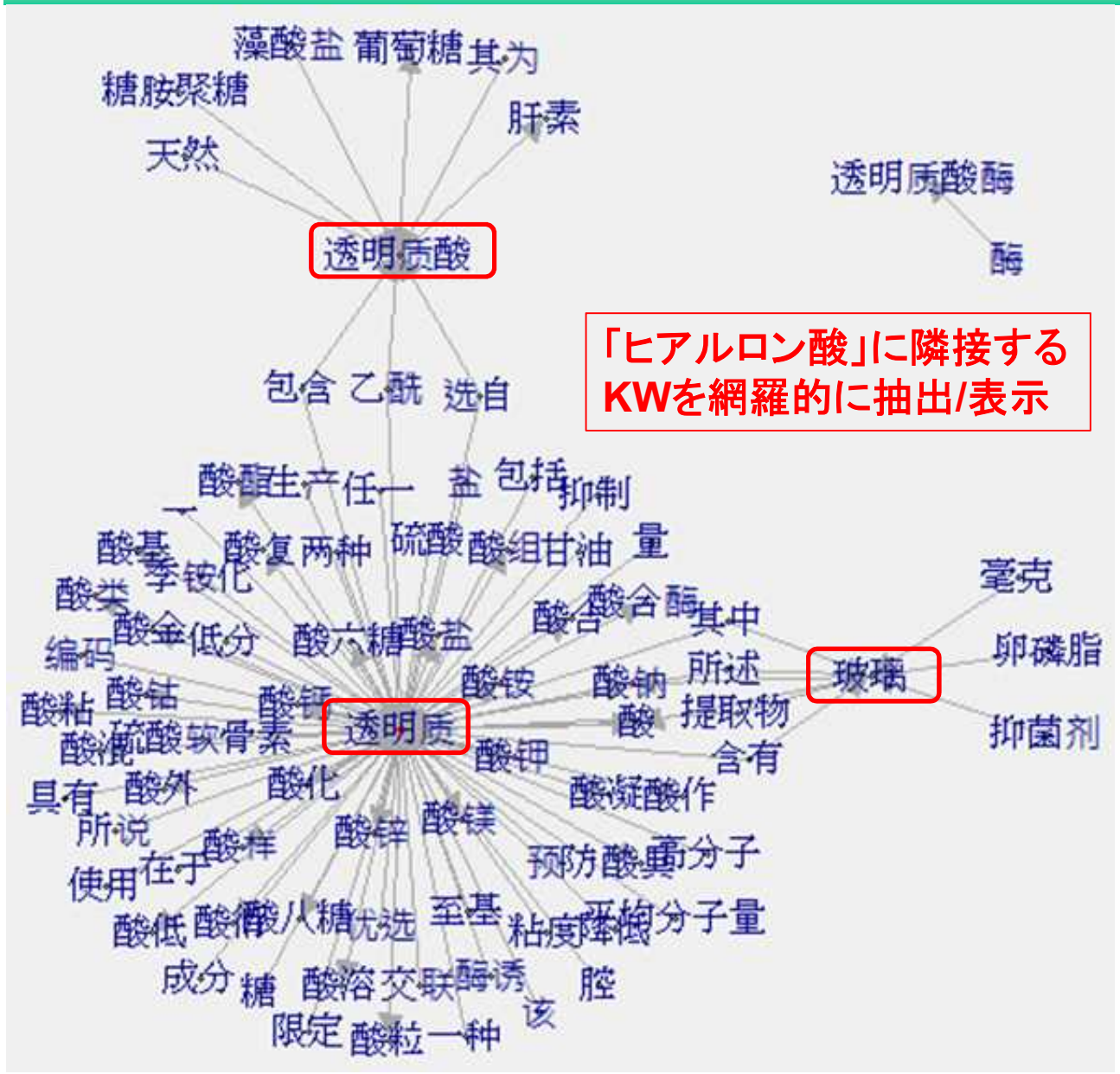
3.如权利要求1或2所述的眼用药物组合物，用于干眼症的治疗或预防。

4.如权利要求3所述的眼用药物组合物，为干眼症滴眼剂。



No.	KW1	KW2	頻度
1	权利要求书	1.一·	1
2	1.一·	眼用	1
3	眼用	药物	1
4	药物	组合	1
5	组合	物	1
6	物	含有	1
7	含有	乙酰化	1
8	乙酰化	透明质	1
9	透明质	酸	1
10	酸	药用	1
11	药用	载体	1
12	2	如权	1
13	如权	1所	1
14	1所	眼	1
15	眼	药物	1
16	药物	组合	1
17	组合	物	1
18	物	上述	1
19	上述	乙酰化	1
20	乙酰化	透明质	1
21	透明质	酸	1
22	酸	平均分子量	1
23	平均分子量	10000	1
24	10000	1000000	1
25	1000000	乙酰	1
26	乙酰	基取	1
27	基取	代数	1
28	代数	2	1
29	2	4	1
30	3	如权	1

「ヒアルロン酸」隣接KWのネットワーク分析



No.	KW1	KW2	頻度
1	透明质	酸	449
2	特征	在于	281
3	组合	物	270
4	其中	所述	268
5	根据	权利要求	258
6	透明质	酸钠	202
7	酸衍	生物	141
8	透明质	酸衍	132
9	所述	透明质	125
10	物	其中	117
11	制备	方法	74
12	在于	所述	71
13	方法	其中	71
14	药物	组合	59
15	所述	化合物	53
16	物	特征	51
17	至少	一·	47
18	任一	限定	47
19	酸	盐	45
20	包含	透明质	45
21	任一	所述	45
22	交联	透明质	45
23	物	包含	44
24	方法	特征	43
25	方法	包括	41
26	其中	透明质	38
27	透明质	酸盐	37
28	用于	治疗	36
29	活性	成分	36
30	海藻	糖	35
2453	玻璃	酸	2

一部抜粋

IPC=A61K31/728(ヒアルロン酸)の検索結果200件（請求項）の上位100組

CNIPRの類似検索検討

①公開番号CN1796780を入力して検索

CN1796780の類似検索

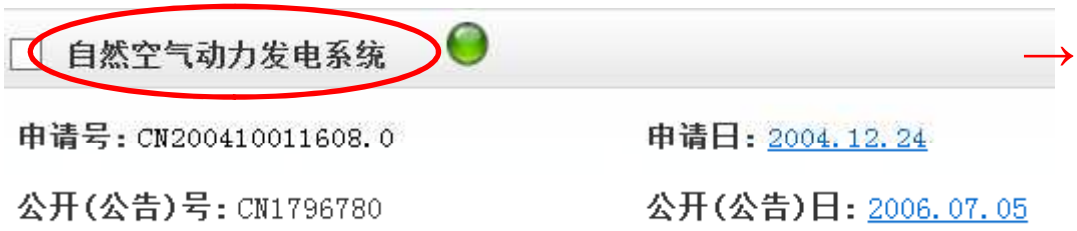
	Search引例
カテゴリーX:	CN1221855, CN1651759 ,CN1619143
カテゴリーY:	CN1261128, CN1405448, CN2479242, CN1257160
カテゴリーA:	CN1454292

CNIPR
新規性検索**115位**

参考：Orbit類似検索**91位**（英語）

PCTサーチ引例に注目して類似検索の性能を評価

②タイトルをクリック



→もっと適合率を向上できないか？

類似性検索
新規性検索
侵害性検索

③類似検索

CNIPRの類似検索結果

相関度を取得

類似性検索：1758件

新規性検索：　225件

侵害性検索：1533件

類似性検索＝新規性検索+侵害性検索

1758件　　225件　　1533件

・新規性検索と侵害性検索は重複なし

・類似検索対象のCN1796780Aの出願日との関係（前後）で振り分けている

CNIPRの類似性、新規性、侵害性検索の相関度

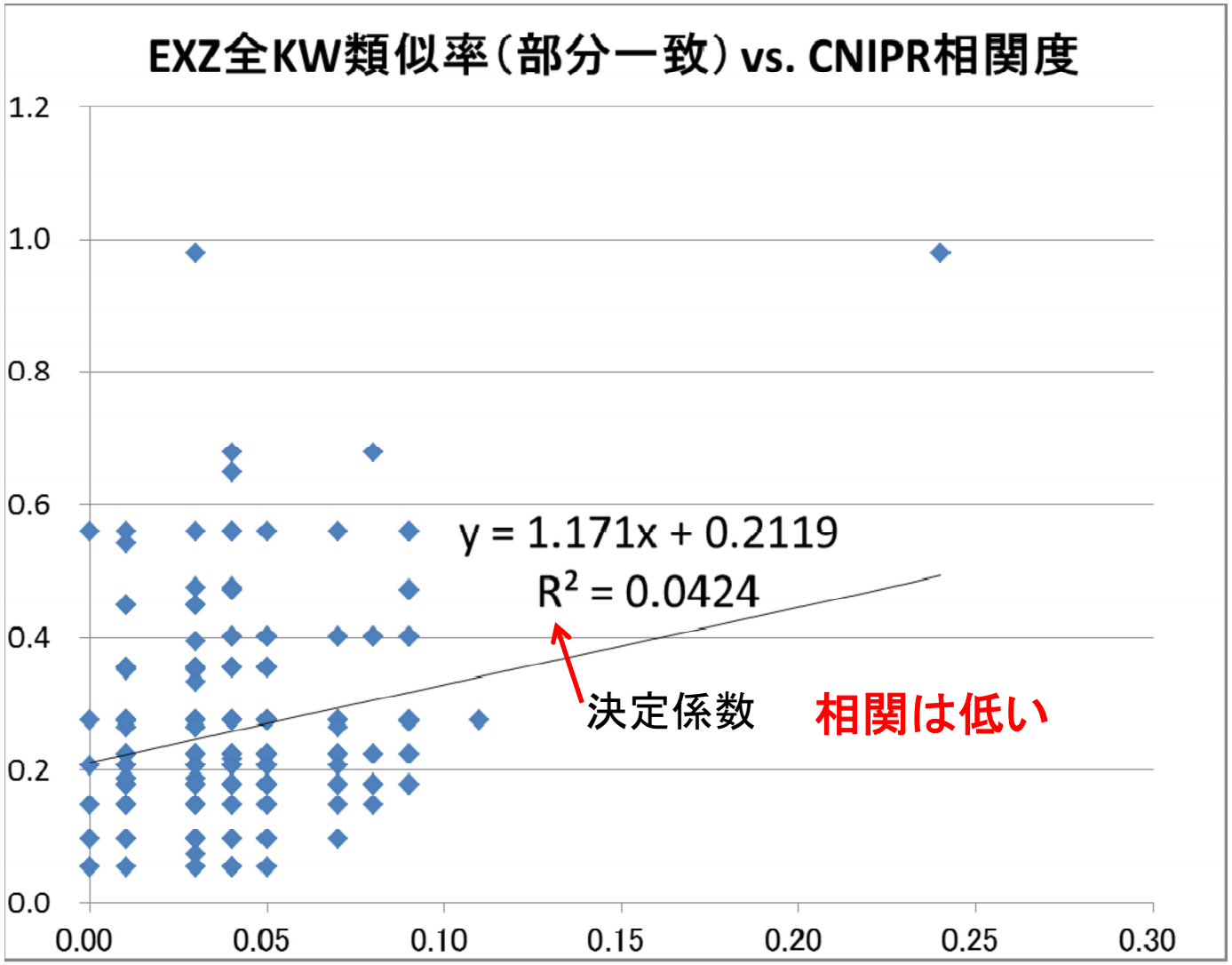
対象：CN1796780A（本願）

	類似性			新規性			侵害性		
CNIPR類似検索	類似性相・度	申请号	公・（公告)号	新規性相・度	申请号	公・（公告)号	侵害性相・度	申请号	公・（公告)号
No.	●	CN200410011608.0	CN1796780	●	CN200410011608.0	CN1796780	●	CN200410011608.0	CN1796780
1	0.98	CN201020127846.9	CN201635943U	0.98	CN97217354.4	CN2290143	0.98	CN201020127846.9	CN201635943U
2	0.65	CN201120339935.4	CN202280578U	0.68	CN200410000013.5	CN1556352	0.65	CN201120339935.4	CN202280578U
3	0.475	CN201120220790.6	CN202152715U	0.68	CN200410012962.5	CN1562660	0.475	CN201120220790.6	CN202152715U
4	0.475	CN201010042760.0	CN102128141A	0.559	CN03246497.5	CN2620838	0.475	CN201010042760.0	CN102128141A
5	0.472	CN201110038948.2	CN102166968A	0.559	CN00131220.0	CN1357466	0.472	CN201110038948.2	CN102166968A
6	0.472	CN201210115179.6	CN102642464A	0.559	CN200410064349.8	CN1609446	0.472	CN201210115179.6	CN102642464A
7	0.472	CN201220166334.2	CN202847436U	0.559	CN01262780.1	CN2502774	0.472	CN201220166334.2	CN202847436U
8	0.402	CN200510012122.3	CN1710741	0.559	CN200420105656.1	CN2753890	0.402	CN200510012122.3	CN1710741
9	0.402	CN201010251495.7	CN101943033A	0.559	CN01119432.4	CN1388318	0.402	CN201010251495.7	CN101943033A
10	0.402	CN201010554365.0	CN102477945A	0.559	CN87207428	CN87207428	0.402	CN201010554365.0	CN102477945A
11	0.402	CN201110291248.4	CN102545702A	0.559	CN01132439.2	CN1342583	0.402	CN201110291248.4	CN102545702A
12	0.402	CN200510058968.0	CN1837609	0.559	CN00123658.X	CN1339865	0.402	CN200510058968.0	CN1837609
13	0.402	CN201120366960.1	CN202300881U	0.543	CN01230658.4	CN2497033	0.402	CN201120366960.1	CN202300881U
14	0.402	CN201110343105.3	CN102434358A	0.449	CN200410059131.3	CN1617431	0.402	CN201110343105.3	CN102434358A
15	0.402	CN200920277338.6	CN201582058U	0.449	CN00130825.4	CN1310290	0.402	CN200920277338.6	CN201582058U
16	0.402	CN200920141664.4	CN201354717	0.449	CN01130082.5	CN1427156	0.402	CN200920141664.4	CN201354717
17	0.402	CN200820128764.9	CN201301779	0.449	CN02285482.7	CN2583416	0.402	CN200820128764.9	CN201301779
18	0.402	CN200710103565.2	CN101050726	0.449	CN00109970.1	CN1336484	0.402	CN200710103565.2	CN101050726
19	0.402	CN200810023929.0	CN101255845	0.449	CN97101903.7	CN1188186	0.402	CN200810023929.0	CN101255845
20	0.402	CN97217354.4	CN2290143	0.449	CN99221454.8	CN2385787	0.394	CN200620050561.3	CN2885666
21	0.394	CN200620050561.3	CN2885666	0.449	CN00246927.8	CN2441983	0.334	CN200510087135.7	CN1710273
22	0.334	CN200510087135.7	CN1710273	0.449	CN03806399.9	CN1642772	0.275	CN200910078109.6	CN101806288A
23	0.275	CN200910078109.6	CN101806288A	0.449	CN88209972.8	CN2032250	0.275	CN201110115210.1	CN102312790A
24	0.275	CN201110115210.1	CN102312790A	0.449	CN01265049.8	CN2552241	0.275	CN201110423176.4	CN102496959A
25	0.275	CN201110423176.4	CN102496959A	0.449	CN98217376.8	CN2387664	0.275	CN201110215445.8	CN102392793A

各上位25件

・新規性検索ではあまり良い結果は得られなかった
・侵害性検索では相関度0.98のCN201635943Uは本願と同じ出願人の類似技術

パテントマップEXZのKW類似率とCNIPRの相関度



CN1796780の
類似検索
新規性検索：224件
侵害性検索：199件
計423件

・類似率計算の基になる抽出KWが異なる
・類似率計算方法がどちらもブラックボックス

適合率向上のための提案手法

提案手法

①ターゲット公報の予備検討

- 発明のポイント抽出
- 重要KW抽出（人手）

②DB検索

- ブーリアン検索
- ダウンロード

③パテントマップEXZへ取り込み

- 重要KWでダミー公報設定
- 類似率でソート

④確認（スクリーニング）

ターゲット公報（本願）：CN1796780A

出願番号	CN200410207838.2
公開番号	CN1796780（WO2006/066502）
出願人	廖意民
発明の名称（中文）	自然空气动力发电系统
要約（中文）	本发明公开了一种自然空气动力发电系统，其包括一具有入气口部分和出气口部分的管身密封的管道，所述管身内设有气轮发电机，所述入气口部分和出气口部分之间具有产生气流足以驱动气轮发电机的气压差。本发明的管道，沿着建筑物的高度方向或环境地势敷设，它不须要实施难度极高的烟囱或不可改变的深井等建筑，因此大大地降低了建筑成本；整个系统可以利用大部分现有的高层建筑或随自然环境的地势而灵活地附加搭建，也可在需要时拆卸搬迁；本发明利用自然的空气动力发电，节约能源，还可抽除高层建筑物底层的停车场、隧道、工厂等的废气、废热，推动环保。应用于机场等大型设施并可减少由于热气流造成的危险，变害为利。
請求項（中文）	1.一种自然空气动力发电系统，其特征在于：包括一具有入气口部分和出气口部分的管身密封的管道，所述管身内设有气轮发电机，所述入气口部分和出气口部分之间具有产生气流驱动气轮发电机运转气流的气压差。 2.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的管身随所依附的地势或建筑物形状而起伏，中途可由两条或两条以上的支管组成主管道。 3.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的入气口部分设置于具有高气压的低位，出气口部分设置于具有低气压的高位，两者间具有产生驱动气轮发电机运转气流的气压差。 4.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述的气轮发电机装置有一台或一台以上；电能的输出电缆敷设在管道内。 5.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管身安装发电机的位置为加宽机房，该机房包括安装有发电机的主管道、副管道、切换气流途径的管道门及机房门。 6.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的入气口部分可设有两条或两条以上的总截面积大于基本管道的进气支管。 7.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的出气口部分上设有上盖装置。 8.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的入气口部分上设有防尘装置。 9.根据权利要求8所述的自然空气动力发电系统，其特征在于：所述的防尘装置为金属丝防尘网罩。 10.根据权利要求1所述的自然空气动力发电系统，其特征在于：所述管道的入气口部分设置于大厦停车场、酒楼排气管、中央空调散热器等废气、废热源的地方或机场等大型设施的热气流多发区域。

提案手法例　類似率ソートフロー

発明のポイント抽出

主請求項	高度差、または温度差により発生する圧力差に起因して発生する自然対流を利用
	風車により発電
	入口と出口を有する密閉配管内に風車を用いた発電機を複数設置
従属項	発電機部の配管を分岐し（発電機を通る流路と通らない流路）、切換機を取り付ける
	山岳地帯と平野部の高度差（例えば3000m）を利用
	ビル内の空調排熱、レストランの調理排熱等を利用
	竜巻やウインドシアを利用することもできる

重要KW抽出（人手）

	日本語	中国語
A	圧力差,高度差,温度差	压差, 高差, 温差, 高度差, 温度差, 压力差, 大气梯度, 气压差
B	対流,空気流,流動	对流, 气流, 流动, 空气下降流
C	風力,発電,風車,タービン	风力, 发电, 风车, 涡轮
D	配管,ダクト	管,风道,气道
E	入口	入口,进口,进风口,进气口
F	出口	出口,排风口,排口,排气口
G	分岐,切換,切替	分枝,交换,切换,转换
H	建物,ビル	大厦,建筑,大楼,号楼
I	廃熱,排熱	余热,废热
J	竜巻	龙卷风, 旋风
K	ウインドシア	风切变

➡ ダミー公報設定データ

重要KW抽出支援方法として
ネットワーク分析の利用検討

DB検索

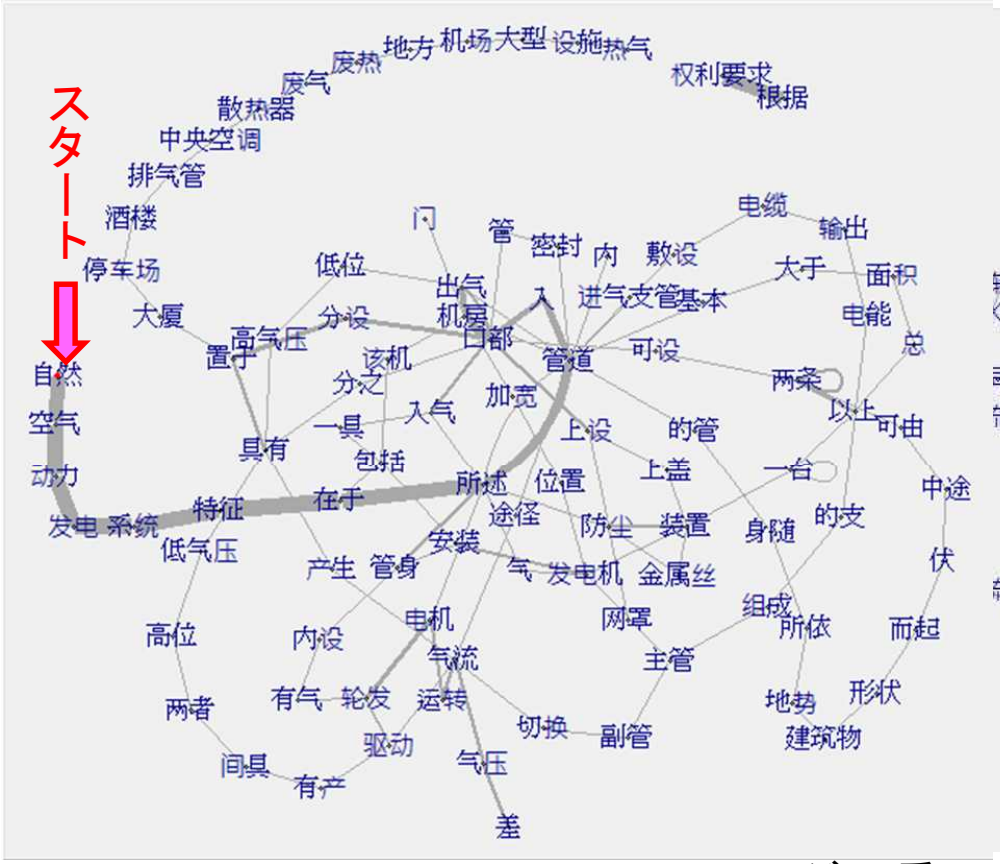
	検索式	集合	特許	実案
IPC	F03G7/04　F03D9/00 F03G6/00	S001		
全文	压差　高差　温差　高度差　温度差　压力差　大气梯度　气压差　对流　气流　流动　空气下降流	S002		
全文	风力　发电　风车　涡轮	S003		
全文	管　风道　气道	S004		
	S001*S002*S003*S004	S005	93	26

検索データベース：HYPAT-i
検索日：2013年7月17日
出願日：2004年12月24日以前

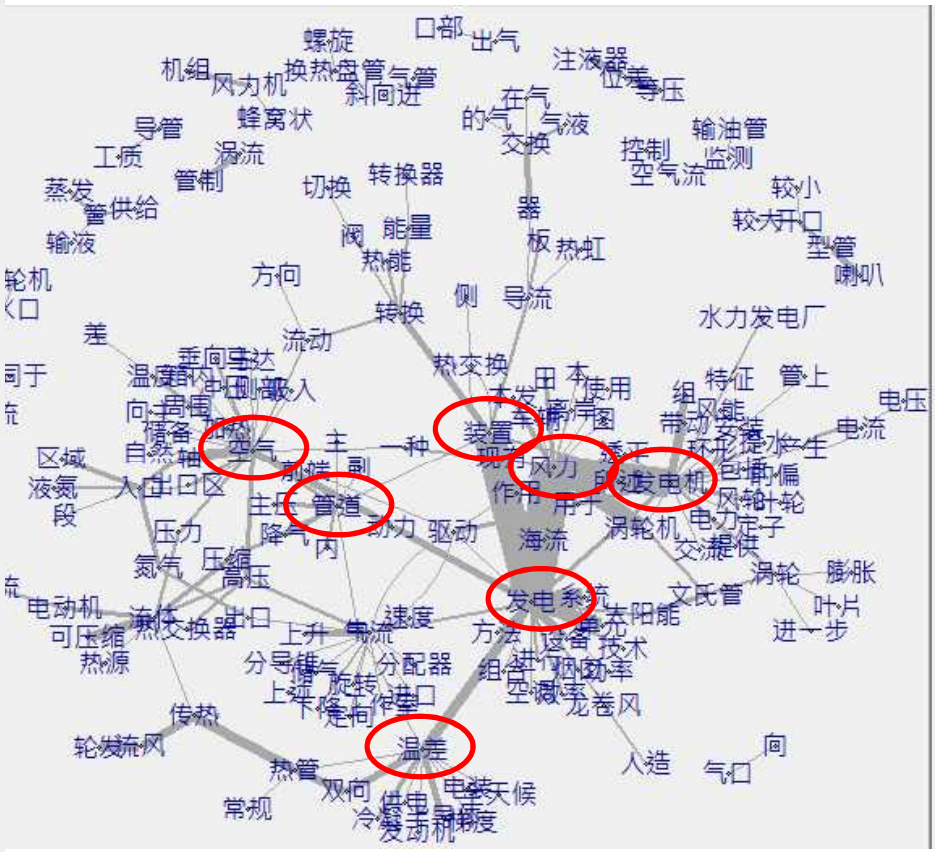
計119件 ➡ 公報全文をEXZへ入力

重要KW抽出支援（ネットワーク分析）

本願CN1796780Aのネットワーク分析



人手抽出の重要KWを含む隣接KWのネットワーク（計119件の公報全文から抽出）

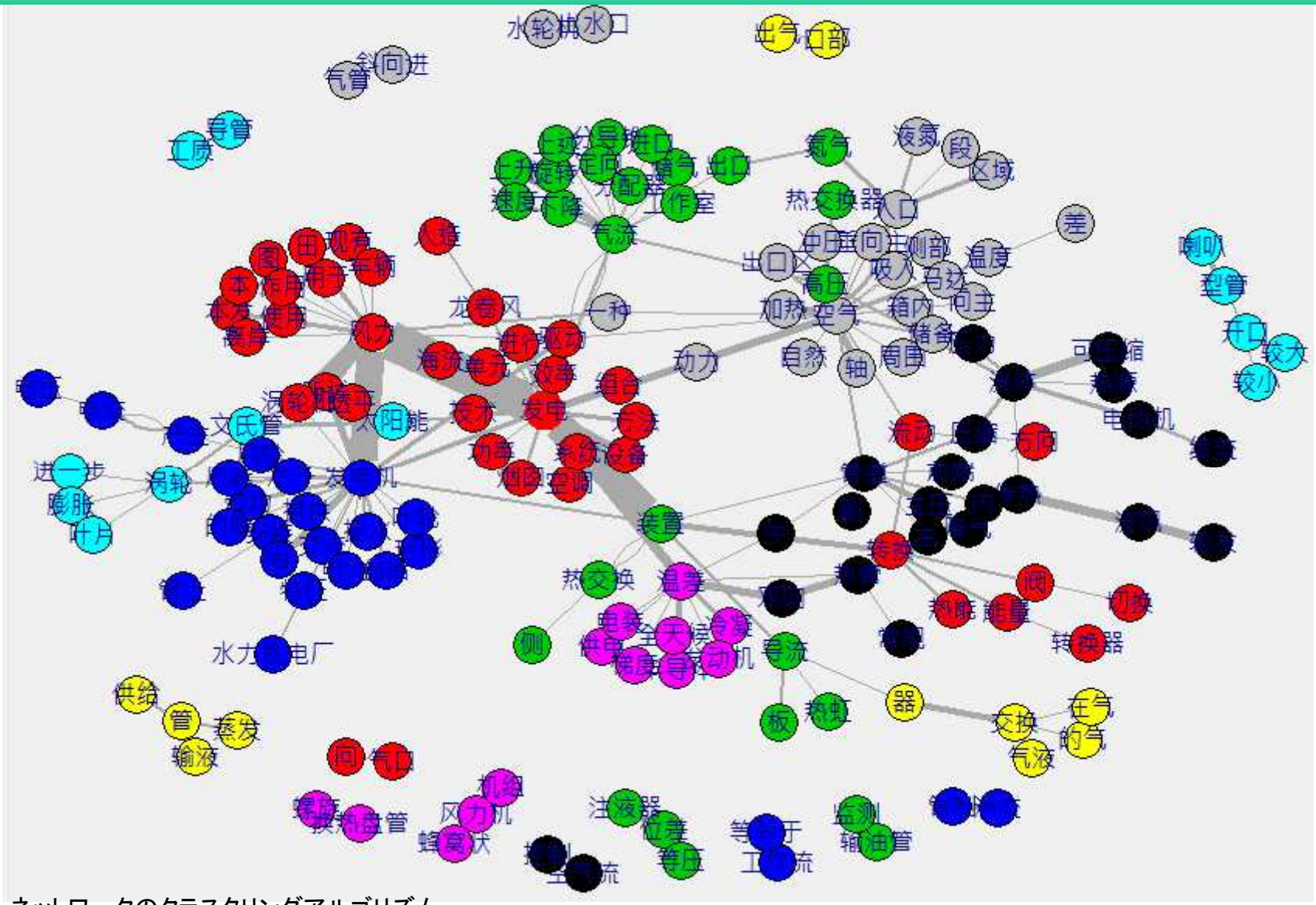


エッジの重み（隣接KW頻度）

①計119件の公報全文
②人手抽出の重要KWを含む文を抽出
③隣接KW対を抽出→ランキング（隣接KW頻度）
④重要KWを含む隣接KW対抽出→ネットワーク分析

探したいKWの隣接KWを抽出
→少し広めのKW（網羅性向上）

隣接KWネットワークをクラスターに分割



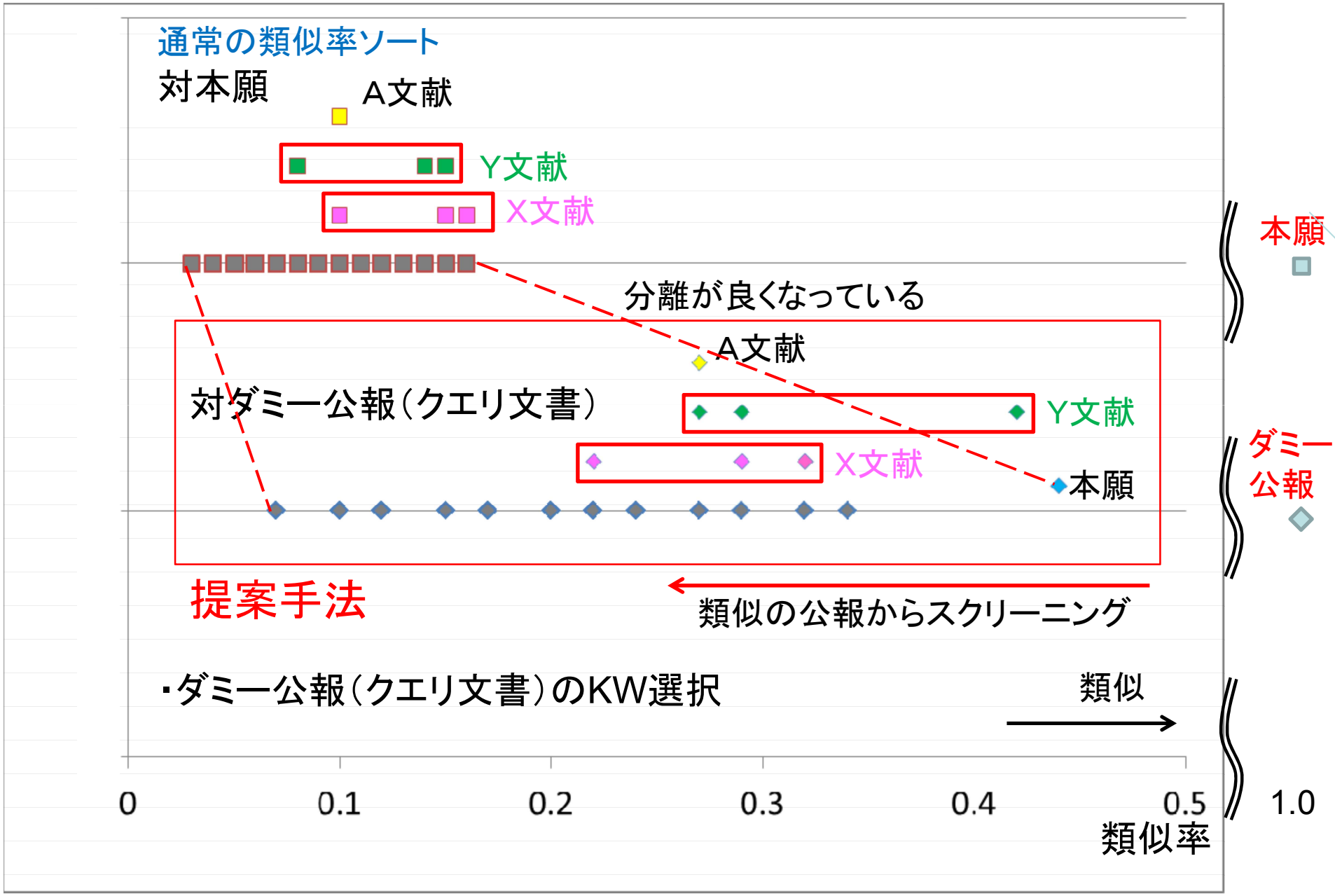
ネットワークのクラスタリングアルゴリズム
スペクトラル最適化法 グラフラプラシアンを使ってQ値が最大となるような分割を探す

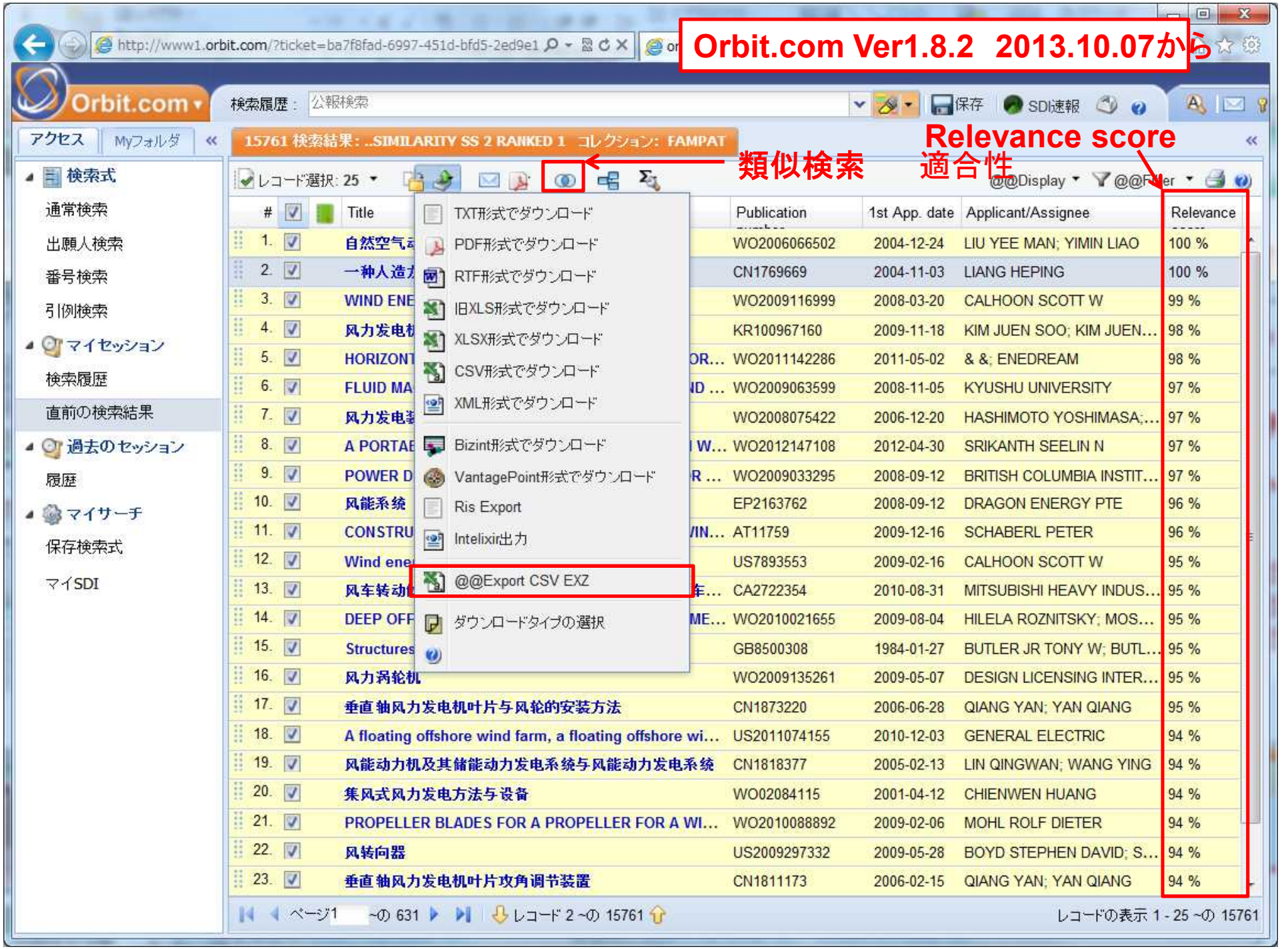
参考：R+igraph問題「友好関係ネットワークから派閥を検出」
http://codeiq.hatenablog.com/entry/2013/06/25/182324

パテントマップEXZの類似率(一次元)ソート結果

	No.	公開・公表・再	出願日	公開・公表・	発明の名称	全出願人	全キーワード類似率(部分一致)
	0				ダミー公報(クエリ文書)		-
本願	1	1796780	2004/12/24	2006/7/5	自然空气动力发电系统	廖意民	44%
Y文献	2	1257160	1998/12/15	2000/6/21	人造龙卷风发电系统	陈玉德;陈玉	42%
	3	2844482	2004/9/2	2006/12/6	温差双向热管传热汽流风轮发电装置	徐宝安	34%
	4	1743666	2004/9/2	2006/3/8	温差双向热管传热汽流风轮发电装置	徐宝安	34%
	5	1997859	2004/4/23	2007/7/11	采用多系统发电和水脱盐的结构和方法	MSC能量私人	32%
X文献	6	1651759	2004/2/3	2005/8/10	利用大气对流层冷空气下降流发电的方法及其	梁和平	32%
	7	1833104	2004/7/7	2006/9/13	风力发电系统、永久磁铁的配置构造以及电/力	宇宙设备	32%
X文献	8	1619143	2003/11/18	2005/5/25	地心引力与大气梯度温差综合发电方法及其装	梁和平	29%
Y文献	9	1261128	1999/8/12	2000/7/26	无向风道高温永恒系统结构	邓百忍;邓伟	29%
	10	1721692	2004/7/16	2006/1/18	空气蓄压结构	林达顺	29%
	11	1215798	1997/3/11	1999/5/5	山坡温室太阳能发电系统	赵松奇	27%
	12	1040082	1988/12/13	1990/2/28	利用环境流体热能的方法	张燕波	27%
	13	1291261	1998/6/30	2001/4/11	风力发电机	艾格·S·奥洛	27%
A文献	14	1454292	2000/10/27	2003/11/5	对流发电方法和装置	阿部俊广	27%
	15	85101085	1985/4/1	#########	活塞式无曲轴液力传动内燃机	陈友年	27%
Y文献	16	2479242	2001/5/18	2002/2/27	风力发电装置	赵佰川;梅长	27%
	17	1429987	2001/12/31	2003/7/16	一·“太阳能全天候温差发电装置系统”	陈绍勇	24%
	18	1192260	1996/6/7	1998/9/2	海洋热能转换系统	奥特克发展	24%
	19	2177815	1993/12/19	1994/9/21	温差能动机	熊福达	24%
	20	1666020	2003/7/11	2005/9/7	具有闭合冷却回路的风力透平	西门子	24%
	21	85106574	1985/8/31	1987/3/18	利用低温和中温源流体的改进型级联发电站	奥马蒂系统	24%
	22	1587690	2004/9/2	2005/3/2	一· 太阳能烟囱发电装置的建造方法	西安交通大学	22%
	23	1509373	2002/4/8	2004/6/30	风动力的水力发电厂及电厂运行的方法	新世界一代	22%
	24	1053108	1990/10/12	1991/7/17	普适温差能发电技术	郑维新	22%
	25	1188526	1996/3/29	1998/7/22	发电和推进装置的螺旋透平	东北大学	22%
	26	1103747	1994/3/11	1995/6/14	太阳能烟囱设备	达雅·兰吉特	22%
X文献	27	1221855	1998/1/1	1999/7/7	山坡太阳能温室造风发电系统	赵松奇	22%
	28	1298061	2000/12/15	2001/6/6	能量转换器	亚历杭德罗	22%
	29	1580546	2004/3/3	2005/2/16	冲气动力风机	李发祥	22%
	30	1773188	2004/11/9	2006/5/17	太阳能风力装置	靳键云;靳少	22%

パテントマップEXZの類似率（一次元）ソート結果




Orbit.com Ver1.8.2 2013.10.07から
Relevance score
適合性
類似検索
Orbit.com
検索履歴：
公報検索
保存
SDI速報
アクセス
Myフォルダ
15761 検索結果：..SIMILARITY SS 2 RANKED 1 コレクション：FAMPAT
検索式
通常検索
出願人検索
番号検索
引例検索
マイセッション
検索履歴
直前の検索結果
過去のセッション
履歴
マイサーチ
保存検索式
マイSDI
レコード選択: 25
@@Display
@@Filter
TXT形式でダウンロード
PDF形式でダウンロード
RTF形式でダウンロード
旧XLS形式でダウンロード
XLSX形式でダウンロード
CSV形式でダウンロード
XML形式でダウンロード
Bizint形式でダウンロード
VantagePoint形式でダウンロード
Ris Export
Intelixir出力
@@Export CSV EXZ
ダウンロードタイプの選択
Title Publication 1st App. date Applicant/Assignee Relevance
1. 自然空气 WO2006066502 2004-12-24 LIU YEE MAN; YIMIN LIAO 100 %
2. 一种人造 CN1769669 2004-11-03 LIANG HEPING 100 %
3. WIND ENE WO2009116999 2008-03-20 CALHOON SCOTT W 99 %
4. 风力发电机 KR100967160 2009-11-18 KIM JUEN SOO; KIM JUEN... 98 %
5. HORIZONT ...OR... WO2011142286 2011-05-02 & &; ENEDREAM 98 %
6. FLUID MA ...ND ... WO2009063599 2008-11-05 KYUSHU UNIVERSITY 97 %
7. 风力发电装 WO2008075422 2006-12-20 HASHIMOTO YOSHIMASA;... 97 %
8. A PORTAB ...W... WO2012147108 2012-04-30 SRIKANTH SEELIN N 97 %
9. POWER D ...R ... WO2009033295 2008-09-12 BRITISH COLUMBIA INSTIT... 97 %
10. 风能系统 EP2163762 2008-09-12 DRAGON ENERGY PTE 96 %
11. CONSTRU ...IN... AT11759 2009-12-16 SCHABERL PETER 96 %
12. Wind ene US7893553 2009-02-16 CALHOON SCOTT W 95 %
13. 风车转动 ...车... CA2722354 2010-08-31 MITSUBISHI HEAVY INDUS... 95 %
14. DEEP OFF ...ME... WO2010021655 2009-08-04 HILELA ROZNITSKY; MOS... 95 %
15. Structures GB8500308 1984-01-27 BUTLER JR TONY W; BUTL... 95 %
16. 风力涡轮机 WO2009135261 2009-05-07 DESIGN LICENSING INTER... 95 %
17. 垂直轴风力发电机叶片与风轮的安装方法 CN1873220 2006-06-28 QIANG YAN; YAN QIANG 95 %
18. A floating offshore wind farm, a floating offshore wi... US2011074155 2010-12-03 GENERAL ELECTRIC 94 %
19. 风能动力机及其储能动力发电系统与风能动力发电系统 CN1818377 2005-02-13 LIN QINGWAN; WANG YING 94 %
20. 集风式风力发电方法与设备 WO02084115 2001-04-12 CHIENWEN HUANG 94 %
21. PROPELLER BLADES FOR A PROPELLER FOR A WI... WO2010088892 2009-02-06 MOHL ROLF DIETER 94 %
22. 风转向器 US2009297332 2009-05-28 BOYD STEPHEN DAVID; S... 94 %
23. 垂直轴风力发电机叶片攻角调节装置 CN1811173 2006-02-15 QIANG YAN; YAN QIANG 94 %
ページ1 ~の 631
レコード 2 ~の 15761
レコードの表示 1 - 25 ~の 15761

文書間相互類似度計算（自作VB.Netプログラム）

VB 2008

推奨

①非類似度（距離）マトリックス計算（2次元）
②類似度ソート用（1次元） ③統計出力

特徴

- 全文書の特徴語、重要度をメモリ上に保持
- 文書間の共通語の抽出にハッシュを使用
- 正規表現によるノイズ除去機能

文書間相互類似度の組み合わせ数

$$\frac{n \times (n-1)}{2}$$

1000件の文書場合

$$\frac{1000 \times 999}{2} = 499500$$

重み付け手法と類似度計算方法

	重み付け	類似度計算方法
1	2値	余弦(Cosine)係数
2	2値	ダイス(Dice)係数
3	2値	ジャカール(Jaccard)係数
4	2値	重複(Overlap)係数
5	2値	単純一致c/a
6	2値	単純一致c/b
推奨 7	重み	余弦(Cosine)係数
8	重み	ダイス(Dice)係数
9	重み	ジャカール(Jaccard)係数
10	重み	単純重み付き

参考　2値データに対する種々の距離（非類似度）

No.	名称	定義	定義域	種別
1	Jaccard 係数	$\frac{a}{a+b+c}$	[0, 1]	A
2	Dice 係数	$\frac{2a}{2a+b+c}$	[0, 1]	A
3	Russell-Rao 係数	$\frac{a}{n}$	[0, 1]	A
4	Sokal-Sneath 係数 (1)	$\frac{a}{a+2b+2c}$	[0, 1]	A
5	Kulczynski 係数 (1)	$\frac{a}{b+c}$	[0, 1]	A
6	Simple Matching 係数	$\frac{a+d}{n}$	[0, 1]	A
7	Hamann 係数	$\frac{(a+d)-(b+c)}{n}$	[0, 1]	A
8	Sokal-Sneath 係数 (2)	$\frac{2a+2d}{a+d+n}$	[0, 1]	A
9	Roger-Tanimoto 係数	$\frac{a+d}{b+c+n}$	[0, 1]	A
10	Sokal-Sneath 係数 (3)	$\frac{a+d}{b+c}$	[0, 1]	A
11	Baroni-Urbani-Buser 係数	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	[0, 1]	A
12	Ochiai(Cosine) 係数	$\frac{a}{\sqrt{(a+b)(c+d)}}$	[0, 1]	A
13	Kulczynski 係数 (2)	$\frac{a(2a+b+c)/2}{(a+b)(a+c)}$	[0, 1]	A
14	Fobes 係数	$\frac{an}{(a+b)(a+c)}$	[0, 1]	A
15	Fossum 係数	$\frac{n(a-1/2)^2}{(a+b)(a+c)}$	[0, 1]	A
16	Simpson 係数	$\frac{a}{\min(a+b, a+c)}$	[0, 1]	A
17	Peason 係数	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[−1, 1]	C
18	Yule 係数	$\frac{ad-bc}{ad+bc}$	[−1, 1]	C
19	McConnaughey 係数	$\frac{a^2-bc}{(a+b)(a+c)}$	[−1, 1]	C
20	Stiles 係数	$\log_{10}\frac{n(\lvert ad-bc\rvert-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$	[0, 1]	C
21	Dennis 係数	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	[−1, 1]	C
22	Mean Manhattan 係数	$\frac{b+c}{n}$	[0, 1]	D

参考文献
中村永友
多次元データ解析法
p196クラスター分析法

風力発電のキーワード解析（Indexと基本統計）

解析対象：風力発電（提案手法例）**119件公報全文**
キーワード抽出方法：IKAnalyzerNetによる分詞
キーワード抽出数：**24431ワード（Term）**

一部抜粋

Term頻度（TF）降順

No.	Term	Term頻度	平均	標準偏差	Min.	Max.	文書頻度	IDF	TF*IDF	文書
1	一	7328	61.58	61.27	2	286	119	1.00	7328.0	CN101001028|CN101027479|CN101415939
2	发电	5423	47.99	45.33	1	219	113	1.05	5703.6	CN101001028|CN101027479|CN101415939
3	装置	3949	35.58	38.77	1	162	111	1.07	4223.8	CN101001028|CN101027479|CN101415939
4	所述	3216	29.24	49.37	1	342	110	1.08	3468.9	CN101001028|CN101027479|CN101415939
5	图	3166	28.27	43.7	1	238	112	1.06	3357.9	CN101001028|CN101027479|CN101415939
6	电机	2581	25.81	32.15	1	194	100	1.17	3030.0	CN101001028|CN101027479|CN101415939
7	中	2469	21.66	26.04	1	159	114	1.04	2575.0	CN101001028|CN101027479|CN101415939
8	发电机	2460	25.63	30.25	1	165	96	1.21	2988.4	CN101001028|CN101027479|CN101415939
9	空气	2394	28.16	51.84	1	371	85	1.34	3199.5	CN101001028|CN101027479|CN101415939
10	风力	2365	29.94	37.89	1	205	79	1.41	3333.9	CN101001028|CN101027479|CN101415939

文書頻度（DF）降順

No.	Term	Term頻度	平均	標準偏差	Min.	Max.	文書頻度	IDF	TF*IDF	文書
1	一	7328	61.58	61.27	2	286	119	1.00	7328.0	CN101001028|CN101027479|CN101415939
2	说明	470	3.95	2.69	1	20	119	1.00	470.0	CN101001028|CN101027479|CN101415939
3	本	2033	17.23	13.48	2	90	118	1.01	2050.2	CN101001028|CN101027479|CN101415939
4	附图	522	4.46	11.03	1	86	117	1.02	530.9	CN101001028|CN101027479|CN101415939
5	利用	1099	9.56	7.97	1	35	115	1.03	1136.6	CN101001028|CN101027479|CN101415939
6	中	2469	21.66	26.04	1	159	114	1.04	2575.0	CN101001028|CN101027479|CN101415939
7	说明书	243	2.13	0.64	1	7	114	1.04	253.4	CN1040082|CN1053108|CN1057715|CN107
8	发电	5423	47.99	45.33	1	219	113	1.05	5703.6	CN101001028|CN101027479|CN101415939
9	图	3166	28.27	43.7	1	238	112	1.06	3357.9	CN101001028|CN101027479|CN101415939
10	装置	3949	35.58	38.77	1	162	111	1.07	4223.8	CN101001028|CN101027479|CN101415939

TF*IDF降順

No.	Term	Term頻度T	平均	標準偏差	Min.	Max.	文書頻度	IDF	TF*IDF	文書
1	一	7328	61.58	61.27	2	286	119	1.00	7328.0	CN101001028|CN101027479|CN101415939
2	发电	5423	47.99	45.33	1	219	113	1.05	5703.6	CN101001028|CN101027479|CN101415939
3	装置	3949	35.58	38.77	1	162	111	1.07	4223.8	CN101001028|CN101027479|CN101415939
4	所述	3216	29.24	49.37	1	342	110	1.08	3468.9	CN101001028|CN101027479|CN101415939
5	图	3166	28.27	43.7	1	238	112	1.06	3357.9	CN101001028|CN101027479|CN101415939
6	风力	2365	29.94	37.89	1	205	79	1.41	3333.9	CN101001028|CN101027479|CN101415939
7	空气	2394	28.16	51.84	1	371	85	1.34	3199.5	CN101001028|CN101027479|CN101415939
8	叶片	1735	33.37	46.82	1	198	52	1.83	3171.4	CN101027479|CN101415939|CN102161376
9	旋转	1893	30.05	63.77	1	291	63	1.64	3096.9	CN101001028|CN101027479|CN101415939
10	风轮	1225	47.12	56.85	1	220	26	2.52	3088.3	CN101415939|CN102161376|CN1215798|CI

自前のKW辞書

インバーテッド
（転置）ファイル

TF：Term Frequency
（**語彙頻度**）網羅性に関係

DF：Document Frequency
（**文書頻度**）

IDF：Inverse Document Frequency
逆文書頻度の対数
計算式

$IDF(t) = \log_{10}(N / df(t))+1$

特定性に関係

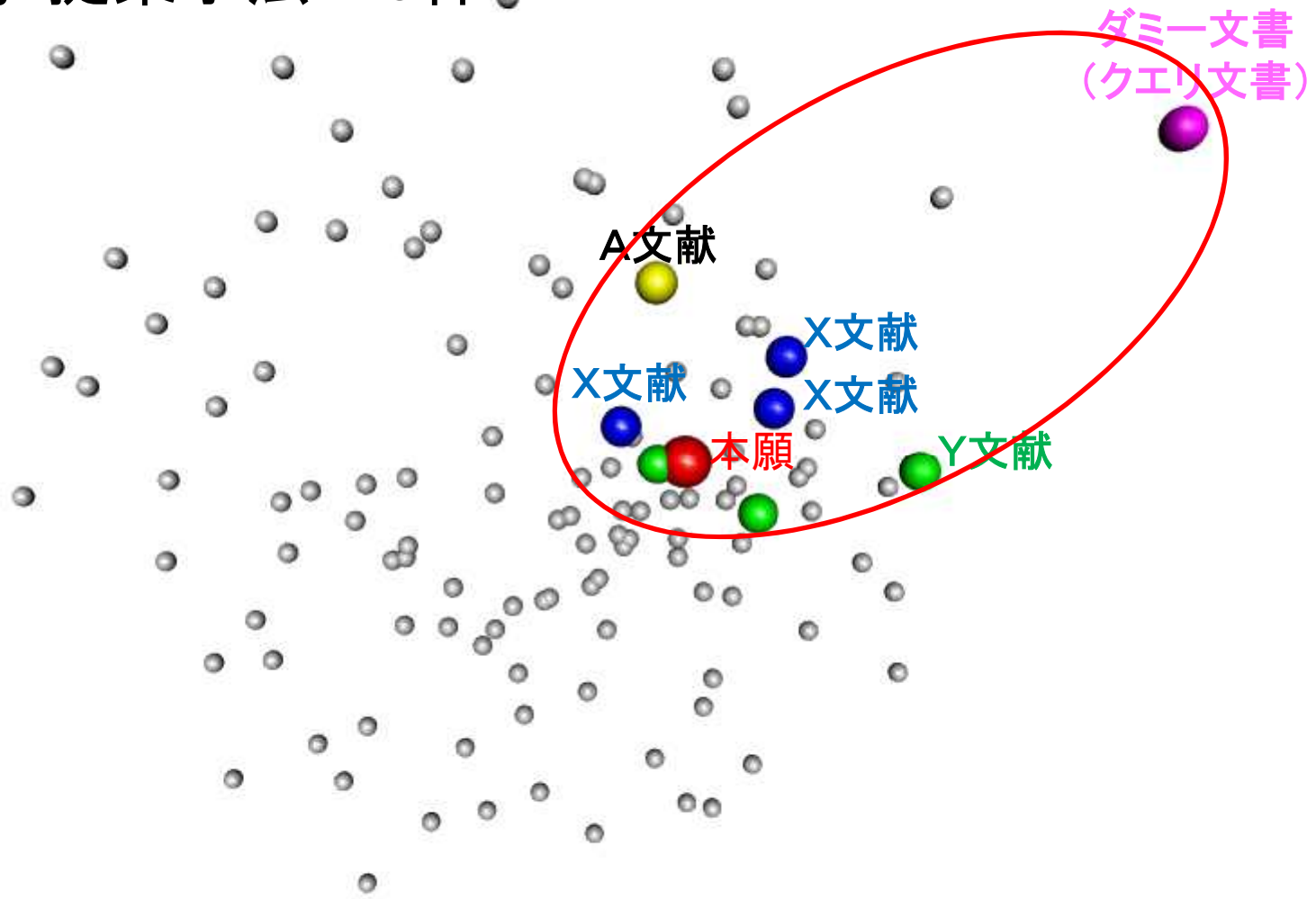
N:全文書数

TF*IDF：TFとIDFの積

適合率重視
→重み付けにもう一工夫

多次元尺度法による公報の2次元表示

風力発電：提案手法119件



類似度計算方法
・コサイン類似度
・重み：TF（ターム頻度）

多次元尺度法による公報の2次元表示

風力発電：CNIPR類似検索423件

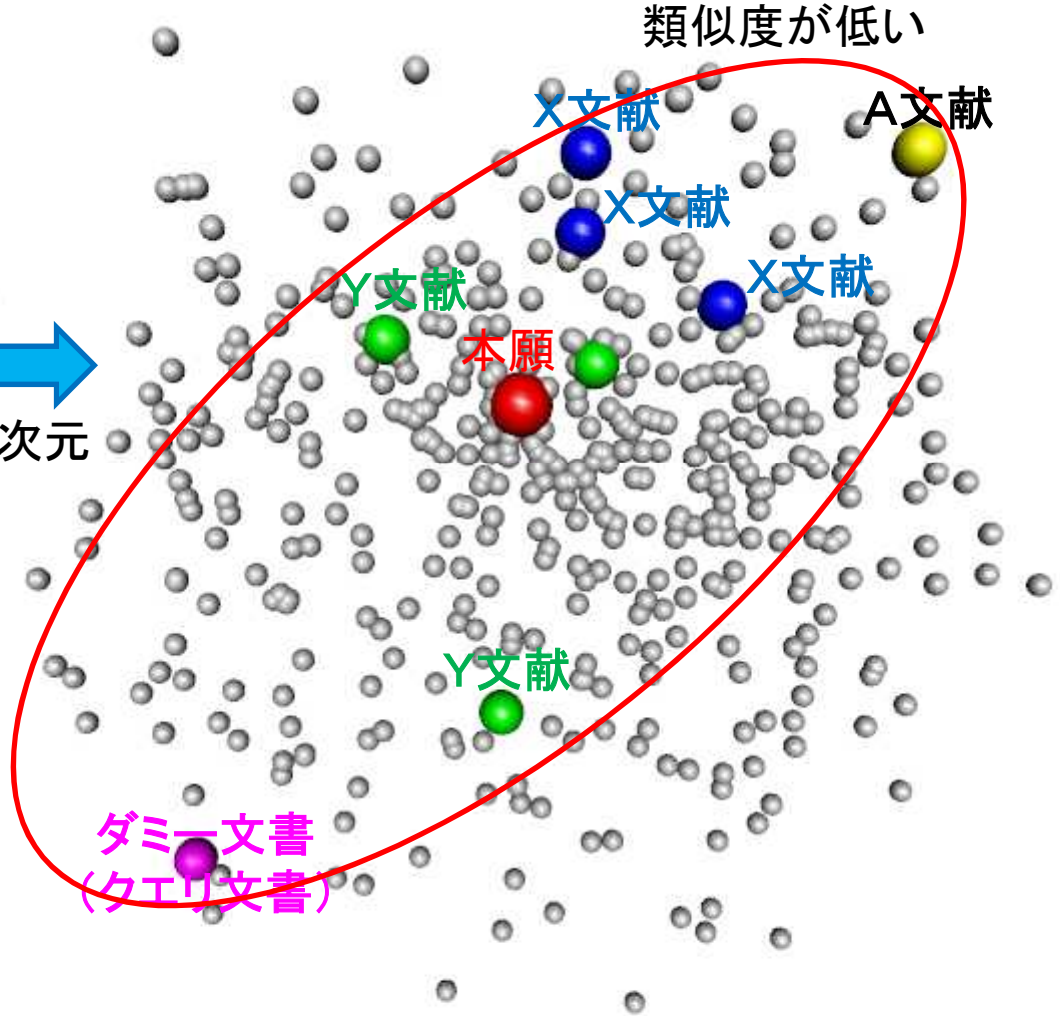
3次元イメージ（ねぎぼうず）
各文献が球状に分布

ダミー文書
（クエリ文書）

注意
視点の位置で見え方が異なる

2次元

元の集合とサーチ引例との
類似度が低い



類似度計算方法
・コサイン類似度
・重み：TF（ターム頻度）

各KWの重み付け、類似度の計算方法等に
より各公報間の相対的位置が変化する

今後の課題

調査用辞書
・専門用語の抽出支援方法（特に化学系の物質名等）
・適合性/網羅性のための特徴語抽出方法の洗練
} →（半）自動化

類似文書抽出
・類似文書抽出のための類似度、重み付け方法
・潜在的意味索引付け(潜在的意味分析、LSA)
→ 同義語抽出、同義語を考慮した類似文書抽出

文書クラスタリング
・ネットワーク分析の文書クラスタリングへの応用

多言語対応
・中国語KW解析
→ 日本語、英語等との共通部分と固有部分を踏まえて多言語へ応用展開

まとめ

中国語KWを用いた特許情報解析に関して、下記の一連の流れを検討した。

①中国語KW抽出
②重要KW抽出
③公報の類似率によるソート
④中国語の概念検索、類似検索
⑤クエリ文書（ダミー文書）の検討、中国特許調査への応用

・重要なKWを選択することで中国特許調査の調査精度、
特に適合率の向上に有用である。

・ターゲット公報を解析した発明の特徴を現す重要KW抽出が
適合率向上のポイントである。

・重要KW抽出にネットワーク分析が有用である。

・類似率ソートにより対象に類似の公報から確認できる。

・文書の類似度は2次元平面上での文書の相互関係の分析にも適用できる。

謝辞

「謝辞」

　最後に、本報告は２０１３年度の「アジア特許情報研究会」のワーキングの一環として報告するものであり、報告者として名前を挙げさせていただいた他に、他テーマリーダーの皆様には様々な協力をしていただきました。

ここに改めて感謝申し上げます。

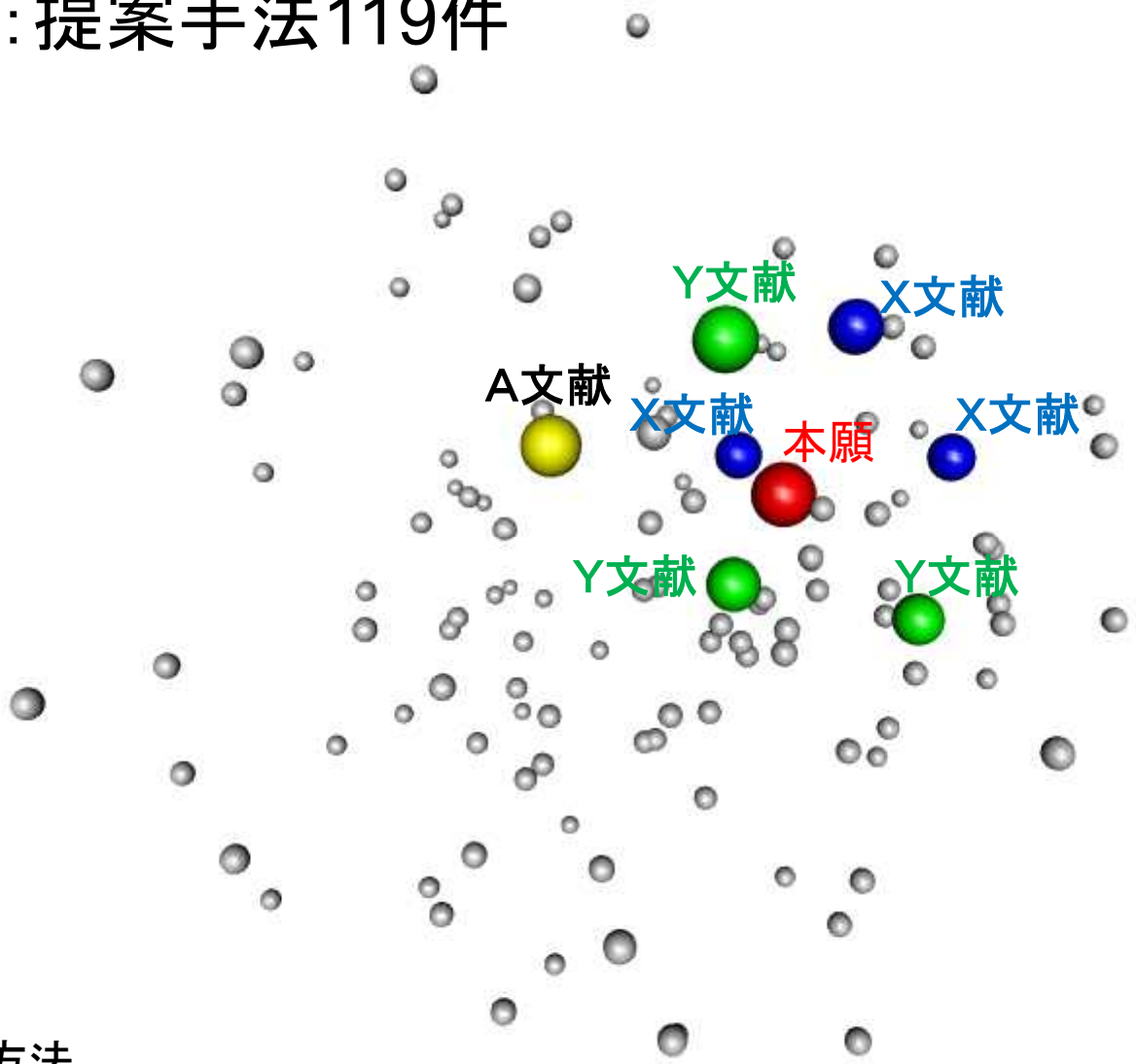
中国特許情報解析チーム一同

ご清聴、ありがとうございました。

補足資料

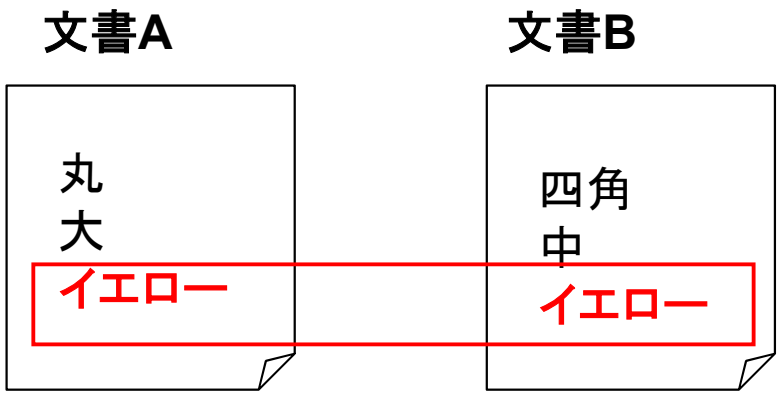
多次元尺度法による公報の3次元表示

風力発電：提案手法119件



類似度計算方法
・コサイン類似度
・重み：TF（ターム頻度）

類似度の計算方法（特徴語の存在を表す2値変数）



a：文書Aのターム数=3
b：文書Bのターム数=3
c：一致したターム数=1

①余弦(Cosine)係数

$$類似度＝\frac{c}{\sqrt{a}\sqrt{b}}\quad\frac{1}{\sqrt{3}\sqrt{3}}＝\frac{1}{3}$$

②ダイス(Dice)係数

$$類似度＝\frac{2\times c}{a+b}\quad\frac{2\times 1}{3+3}＝\frac{2}{6}＝\frac{1}{3}$$

③ジャカール(Jaccard)係数

$$類似度＝\frac{c}{a+b-c}\quad\frac{1}{3+3-1}＝\frac{1}{5}$$

④重複(Overlap)係数

$$類似度＝\frac{c}{\min(a,b)}\quad\frac{1}{3}$$

⑤単純一致

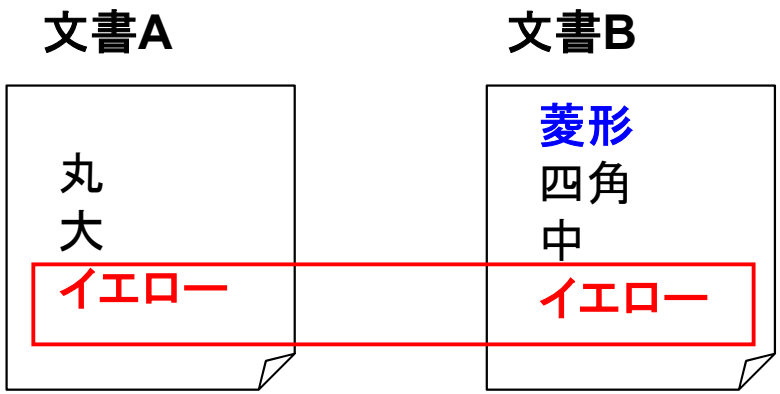
$$類似度＝\frac{c}{a}\quad\frac{1}{3}$$

（パテントマップEXのキーワード類似率）

⑥単純一致

$$類似度＝\frac{c}{b}\quad\frac{1}{3}$$

類似度の計算方法（特徴語の存在を表す2値変数）



a：文書Aのターム数=3
b：文書Bのターム数=4
c：一致したターム数=1

①余弦(Cosine)係数

$$類似度＝\frac{c}{\sqrt{a}\sqrt{b}}\quad \frac{1}{\sqrt{3}\sqrt{4}}\fallingdotseq 0.289$$

②ダイス(Dice)係数

$$類似度＝\frac{2\times c}{a+b}\quad \frac{2\times 1}{3+4}=\frac{2}{7}\fallingdotseq 0.286$$

③ジャカール(Jaccard)係数

$$類似度＝\frac{c}{a+b-c}\quad \frac{1}{3+4-1}=\frac{1}{6}\fallingdotseq 0.167$$

④重複(Overlap)係数

$$類似度＝\frac{c}{\min(a,b)}\quad \frac{1}{3}\fallingdotseq 0.333$$

⑤単純一致

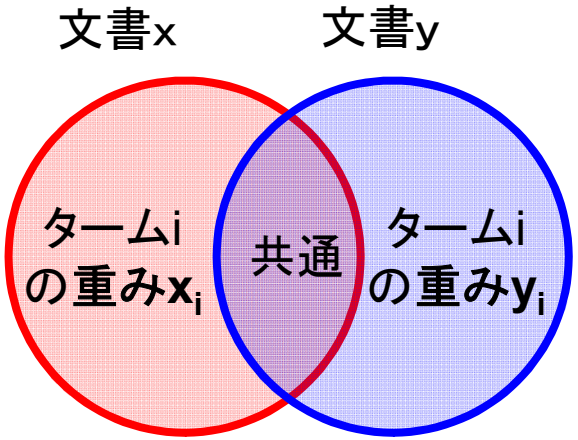
$$類似度＝\frac{c}{a}\quad \frac{1}{3}\fallingdotseq 0.333$$

（パテントマップEXのキーワード類似率）

⑥単純一致

$$類似度＝\frac{c}{b}\quad \frac{1}{4}\fallingdotseq 0.25$$

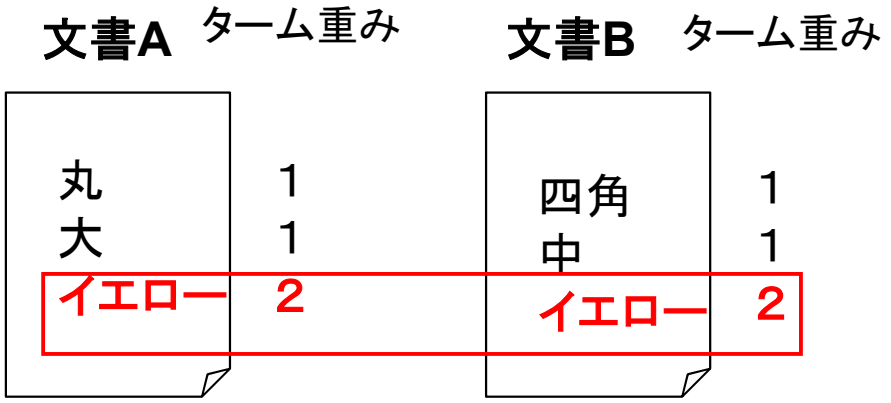
類似度の計算方法（特徴語の重みを使用）



x_i、y_i、文書x、yの特徴語i の**重み**

重み：termmiの重要度、TF・IDF等

計算例



$$\frac{2\cdot 2}{\sqrt{(1^2+1^2+2^2)\times(1^2+1^2+2^2)}}=\frac{4}{6}=\frac{2}{3}$$

①余弦(Cosine)係数

$$\text{類似度}=\frac{\sum x_i\cdot y_i}{\sqrt{\sum x_i^2\times\sum y_i^2}}$$ **0.667**

②ダイス(Dice)係数

$$\text{類似度}=\frac{2\sum x_i\cdot y_i}{\sum x_i^2+\sum y_i^2}$$ **0.667**

③ジャカール(Jaccard)係数

$$\text{類似度}=\frac{\sum x_i\cdot y_i}{\sum x_i^2+\sum y_i^2-\sum x_i\cdot y_i}$$ **0.5**

④’単純重み付き

$$\text{類似度}=\frac{\sum x_i\,(\text{共通})+\sum y_i\,(\text{共通})}{\sum x_i+\sum y_i}$$ **0.5**

参考文献：[8] 徳永健伸. “検索語の重み付け”. 情報検索と言語処理. 東京大学出版会,1999,pp26-33