研究紹介　検索可視化とテキストマイニング

廣川　佐千男　情報基盤研究開発センター

http://tml.cc.kyushu-u.ac.jp/

1.　はじめに

具体的な文書群について検索エンジンを構築し、その文書の特徴と利用目的に応じた検索インターフェースを検討することで、新しい検索エンジンの構築を目指しています。企業情報、特許明細書、学術文献情報など、大量だが限定されたデータの分析法の研究を行っています。アプローチの特色は、単純な手法とユニークな可視です。これまで、キーワードの関連を可視化する概念グラフ、検索結果を多面的に表示するマトリックス検索、クロス表、二重ランク法、制約付ブートストラップなど、新しい分析方式を考案し、論文ならびに特許として発表してきました。

2．研究活動

研究を行ってきた検索手法についてこれまでに 11 件の特許出願を行い、３件が登録されました。その中でも、キーワードの関連を表す概念グラフの実用化研究として、平成 18 から 20 年度まで科学技術振興機構（JST）「独創的シーズ展開事業大学発ベンチャー創出推進」[１]に採択され起業準備をしました。この活動について、平成 19,20 年度に本学から研究・産学連携活動表彰を受けました。その成果として、平成 19 年 9 月には、株式会社 Lafla[２]を設立し、概念グラフを利用した有価証券報告書の分析システムの商用サービスを行っています。昨年一昨年には、諸君の先輩も入社してくれて頑張っています。昨年９月には、九大で先端応用情報科学国際会議 AAI2012 [３]を主催しました。「ユニークであれ」という研究室のモットーを、「居酒屋でバンケット」という形で実践し、国内、国外の参加者にも大いに受けました。ちなみに、今年は AAI2013[４]を松江で８月３１日から９月４日まで開催します。

３．　研究課題の事例

図１は、概念グラフ（特許 2006-082433）[4]を応用した、有報 Lenz という名前でサービス[5]の図です。３つの企業、あるいは一つの３期の企業情報に現れる特徴語の関連をネットワークとして可視化したものです。

図２は、特許明細書を二つの観点で分析するマトリックス検索（特許 4667889）[6]での検索結果です。エネルギーで検索した結果を、課題を縦軸、特許分類番号を横軸にして分類したものです。このような図はパテントマップと呼ばれ、競合企業の分析や新しい研究テーマの検討に必須のツールです。従来はこのようなマップは人手で作られているのですが、本手法はこの自動生成を実現します。

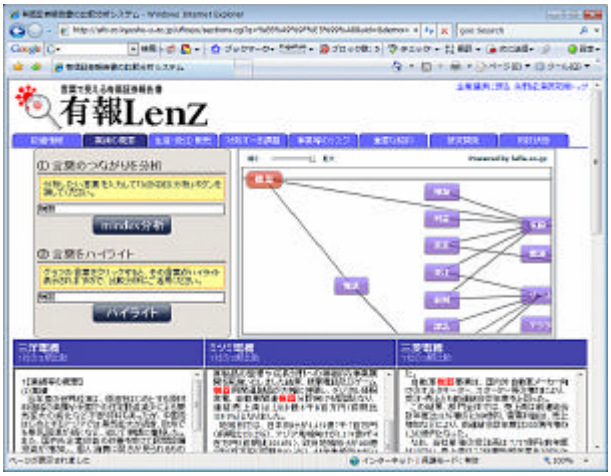

図１　概念グラフ

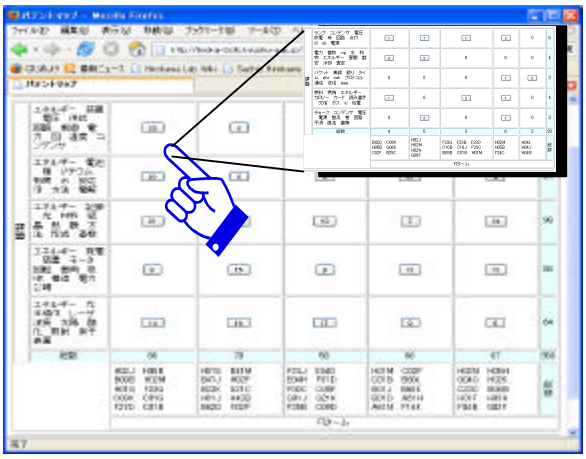

図２　マトリックス検索

図３　クロス表による文献検索

図３は、同様な手法を文献検索に適用したものです。出版年、研究者、研究グループなどを分析軸とすることで、研究調査が効率化できます[7]。

ラーメン 130　しっかり(15) こってり(5) まろやか(2) しっとり(2) ふんわり(2)
うどん 65　しっかり(19) まろやか(15) ぎっしり(14) もっさり(13) ぶるぶる(13)

		ラーメン→			
		1	2	3	
うどん↓	1	しっかり(15, 19)			
	2			まろやか(2, 15)	
	3				ぎっしり(0, 14)
	4				ぶるぶる(0, 13) もっさり(0, 13)
			こってり(5, 0)	しっとり(2, 0) ふんわり(2, 0)	

図４　ダブルランク法

図４も二次元表示は同じですが、表中のセルには件数だけでなく、特徴的キーワードを表示しています。この図は、観光関連ブログデータを対象として、ラーメンとうどんで検索し、それぞれの検索結果に現れる特徴的なオノマトペをダブルランク法 [8,9,10]で表示したものです。単純なキーワードでの分析でなく、感性的な分析を実現しようと思っています。

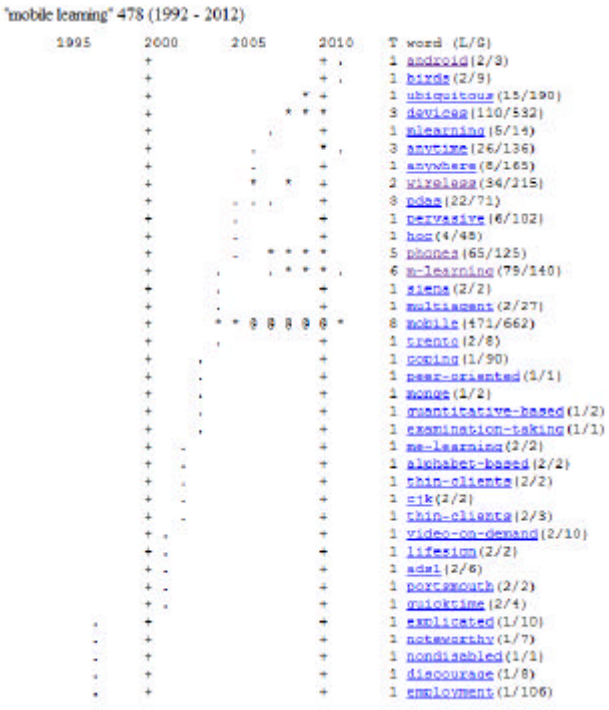

図５　MilkyWay

図５は、時間軸に注目しキーワードの変化を捉えたMilkyWay システム７）です。1992 年から 2012 年に発表された e-learning 関連の文献情報を対象に、時代による研究動向を天の川のように表示することで、個別の論文を見ていてもわからない概観を捕まえようとしています。

システム作りの体力が必要ですが、実データを対象に自分のアイデアで、誰も見たことがないような分析結果を出してみましょう。できたら実用化、企業化までやる心意気のある人を歓迎です。

文献

[1]　平成 18 年度 JST 大学発ベンチャー創出推進「データマップ法と概念グラフによる次世代検索エンジンの研究開発」平成 20 年度終了課題事後評価報告書 http://www.jst.go.jp/tt/uventure/h18hyouka/04-08.html

[2] 株式会社 Lafla, http://www.lafla.co.jp

[3] 先端応用情報科学国際会議 AAI2012, http://aai2012.iaiai.org/

[4] 廣川　佐千男，下司　義寛，和多　太樹，特許 2006-082433, データマップ作成サーバ、データマップ作成方法、およびデータマップ作成プログラム, 2006

[5]　有報告 Lenz, www.yano.co.jp/ufolenz/

[6] 廣川 佐千男，関 隆宏，山田 泰寛，特許 4667889,データマップ作成サーバ、データマップ作成方法、およびデータマップ作成プログラム,2005

[7]　B. Flanagan, C. Yin, S. Hirokawa, H.-Y. Sung, G.-J. Hwang, Analysing Research Trends of Mobile

Learning with the Milky Way, Proc. 7th International Conference on Wireless, Mobile and Ubiquitous Technology in Education (WMUTE2012), pp.249-253, 2012

[8] **廣川 佐千男, 御手洗 秀一, 特願** 2011-104418, **情報処理装置、情報処理方法及びプログラム**,2011

[9] **廣川 佐千男, 二つの観点に基く検索結果の分析方法** Double Rank **について,検索,第** 1 **回テキストマイニング・シンポジウム,電子情報通信学会技術研究報告**. NLC, **言語理解とコミュニケーション** 111(119), pp.61-65, 2011

[10] S. Hirokawa, J. Zeng, Comparison of Tourism Data using Double Ranking, Proc. SSNE, International Workshop on Innovative Tourism Informatics pp.67-72,2011