

筑波大学大学院博士課程

システム情報工学研究科修士論文

文書検索における検索クエリの作成を支援する
視覚的インタフェースの開発

蘇 紹華

修士（工学）

（コンピュータサイエンス専攻）

指導教員 三末和男

2015年3月

概要

特許文献の検索や法律における判例の検索など、プロフェッショナルの文書検索では再現率の高い検索結果が求められる。そのような要望に対して、まず広いキーワードで検索する。そして、得られた検索結果をクラスタリングする。最後に、マイナスキーワードを追加することで不要なクラスタを排除するという手法がある。本研究では、そのような手法を効率的に行うための視覚的インタフェースを開発した。そのインタフェースは、クラスタリングされた検索結果を利用してクラスタ間の関連性を可視化し、クラスタの保存あるいは排除に有効なキーワードを提示する。さらに、修正したクエリを瞬時に実行して実行結果をプレビューできる機能を備える。この視覚的インタフェースによって、適切なクエリを構築することが容易になった。

目次

第1章	はじめに	1
1.1	文書検索の目的	1
1.2	文書の収集	1
1.3	文書の収集における問題	2
1.4	本研究の目的とアプローチ	4
1.5	本論文の構成	4
第2章	関連研究	5
2.1	文書の収集手法	5
2.2	検索結果の提示手法	6
第3章	クエリ作成におけるインタフェイス設計の要件	7
3.1	可視化についての要件	7
3.2	ツール操作の要件	9
第4章	MQViz：マイナスキーワード作成を支援するツールの設計と開発	10
4.1	システム構成と操作の流れ	10
4.2	ツール画面の各エリア	12
4.3	視覚的な表現	13
4.3.1	クラスタの視覚的表現	13
4.3.2	クラスタの関連性の視覚的表現	14
4.3.3	キーワードの視覚的表現	14
4.3.4	クエリの実行効果の視覚的表現	17
4.4	操作	17
4.4.1	クラスタの選択	20
4.4.2	モードチェンジ	20
4.4.3	クエリの試行錯誤	21
4.4.4	Fisheye 機能	21
4.4.5	補助的機能	23
4.5	実装	23
第5章	適用例	24
5.1	事例：「セブン-イレブン」についての文書収集	24

5.2	利用データの詳細	24
5.3	クエリセットの構築	24
第 6 章	まとめ	30
	謝辞	31
	参考文献	32

目 次

1.1	クラシックな情報検索モデル	2
1.2	マイナスクエリを利用する時の情報検索モデル. (a) クラシックな情報検索モデルと違って, まず, 広いキーワードでできるだけ多くの文書を収集する. そして, マイナスキーワードを追加することによって不要な文書を排除する. . .	3
3.1	本研究の情報検索モデル. (a) ユーザーがクラスタリングの結果を参考してクエリを構築する. (b) ユーザーは繰り返しクエリを再構築する.	8
4.1	システム構成図	10
4.2	操作の流れ	11
4.3	ツール画面上の5つのエリア	12
4.4	クラスタの視覚的表現	13
4.5	左右非対称のクラスタの描画方法	14
4.6	クラスタの詳細情報	15
4.7	クラスタの関連性の視覚的表現	15
4.8	キーワードの4つの状態	16
4.9	観察モード	18
4.10	削除モード	19
4.11	クエリの実行効果の視覚的表現	20
4.12	クラスタの選択 (緑色)	21
4.13	Fisheye 機能を使って共有キーワードの探索	22
5.1	クラスタを文書数順にソートする	25
5.2	クラスタを1つ1つ閲覧する	26
5.3	緑色を重ねたクラスタはマークしたクラスタ, つまり消したいクラスタ . . .	27
5.4	削除モード	28
5.5	削除モードでマイナスキーワードを選択する	29

第1章 はじめに

1.1 文書検索の目的

文書検索は大量の文書から目的に合致したものを取り出すために利用されている。文書検索の検索性能を評価するには適合率 (precision) と再現率 (recall) の2つの指標が用いられる。適合率は検索結果として得られた集合の中にどれだけ検索に適合した文書を含んでいるかを表す正確性の指標であり、再現率は検索対象としている文書の中で検索結果として適合している文書のうちでどれだけ文書を検索できているかを表す網羅性の指標である。

文書検索の目的によって、2つの指標の一方がより求められている場合が存在する。例えば蔵書や商品を検索する時、欲しいものを明白に知っている場合、高い適合率が求められる。一方で、マーケット調査、マーケティングリサーチをする時、一つのトピックスに関して文書なるべく多く収集したい場合、高い再現率が求められている。2つの指標が高いほど良い検索システムであるが、2つの指標を同時に上げることは容易でない。

1.2 文書の収集

特許文献の検索や法律における判例の検索など、プロフェッショナルの文書検索領域における検索の特徴として、1) 検索者は一つのタスクに対して大量のクエリセットを繰り返して試す、2) 検索結果の適合率が求められている、3) ブーリアンクエリが一番よく使われている [1]、という点が挙げられる。ブーリアンクエリは簡単に作成できるだけではなく、字面だけで意味が理解できることが使われている原因だと考えられる。また、これらの検索の目的は検索対象に関連する全ての文書の収集である。

クラシックな情報検索モデル [2] を図 1.1 に示す。まず、検索対象に関する検索キーワードを入力して検索する。次に、欲しい情報だけを捉えるために、検索キーワードを変えるか、絞り込みキーワードを追加することになる。しかし、検索結果が絞り込まれすぎて、本来残したい情報も複数の検索キーワードによって排除されることがある。

そこで、以下の手法が考えられる (図 1.2 を参照)。まず、広いキーワードでできるだけ多くの文書を収集する。例えば、コンビニエンスストアの「セブン-イレブン」に関する文書を収集する時、正式名称「セブン-イレブン」を書く文書と略称「セブン」を書く文書があると考えられる。ここで、広く文書を収集できる代表的なキーワードとして「セブン」で検索する。しかし、関係ない文書が引かかるため、マイナスキーワードを追加することによって不要な文書を排除する。「セブン」の検索例だとすると、「ウルトラセブン」など要らない文書

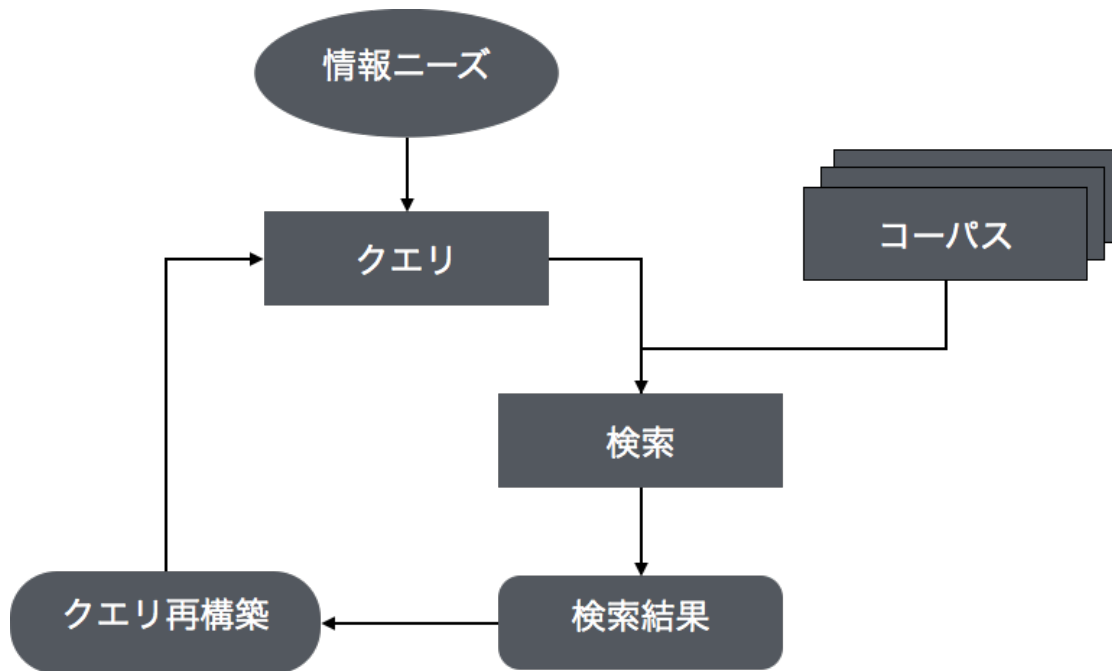


図 1.1: クラシックな情報検索モデル

が引っかかるため、「ウルトラ」をマイナスキーワードに加えるによって、検索結果のノイズを消すことができる。つまり、「一つの検索キーワード+複数のマイナスキーワード」を用いた検索によって、適合率と再現率の高い文書検索が可能だと考えられる。

また、クラスタリングでは大量の文書の概要を提供できる [3] [4] ため、検索結果にクラスタリングを行う。そして、各クラスタの意味を把握した後、要らないクラスタを消すためのクエリセットを作成することで、検索効率を向上させることが可能だと考える。

1.3 文書の収集における問題

しかし、上述の文書収集方法を実際に応用する際には、クエリセット構築とクラスタリングの段階に問題が存在する。

クエリセットを構築する時、クエリとして使えるキーワードを探すために、文書からクエリセットに用いるキーワードを入力する。そうした場合、繰り返して大量のキーワードを検索しなければならず、効率が悪い。

一つのキーワードで検索された大量の文書を閲覧するために、クラスタリングを行う。しかし、クラスタリングの結果によっては、複数のトピックが一つのクラスタに含まれることや複数のクラスタが同じトピックになってしまうことがある。

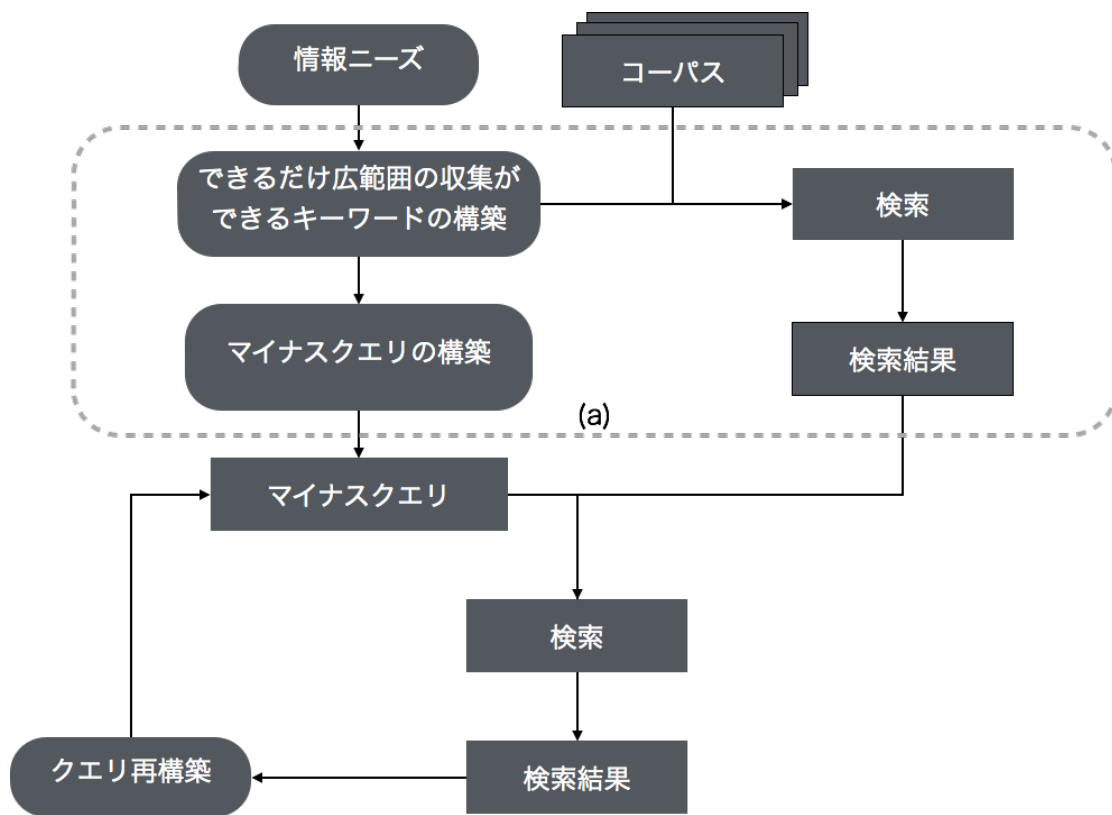


図 1.2: マイナスクエリを利用する時の情報検索モデル. (a) クラシックな情報検索モデルと違って, まず, 広いキーワードでできるだけ多くの文書を収集する. そして, マイナスキーワードを追加することによって不要な文書を排除する.

1.4 本研究の目的とアプローチ

本研究はプロフェッショナルの文書検索領域における文書収集に着目し、文書検索の再現率の向上を目的とし、そのために検索クエリの作成を支援する視覚的表現を開発する。

本研究では、「一つの検索キーワード+複数のマイナスキーワード」のような検索クエリを作成することで文書の収集を行う。このマイナスキーワードの選択を支援するために、プレビューの提示と素早い実行が可能なインタフェースを開発する。また、大量の検索文書を閲覧するためにクラスタを行う。クラスタリング手法として、Latent Dirichlet Allocation(LDA) [5]を採用する。LDAは教師無しで文書集合からトピックの推定できるという特徴を備えているため、文書集合に有効のオーバービューを提供できる [6]。その後、クラスタの重なりを考慮した可視化を行い、クラスタリング結果の精度を検索者に提示する。最後に、クエリ選択の試行錯誤の効率を上げるために、マイナスキーワードとして使えるキーワードを順番付きで提示する。これにより、提示されたキーワードが瞬時にマイナスキーワードとして実行できる。

開発したインタフェースは以下の特徴を備える。

1. マイナスキーワードとして使えるキーワードを提示できる
2. マイナスキーワードを瞬時に実行しながら作成することができる
3. 作成されたマイナスキーワードの実行結果を視覚的に確認できる

このような特徴は、特許情報のサーベイや市場調査などの様々の分野に応用することができる。

1.5 本論文の構成

第二章では、関連研究を紹介する。続く第三章において、MQVizにおける設計要件を定義する。次に、第四章では、MQVizの視覚的表現と操作について紹介する。第五章では、MQVizを用いて具体的なユースケースを紹介する。最後に、第六章で本研究のまとめを述べる。

第2章 関連研究

本研究はプロフェッショナル向けの文書収集に重点を置き、検索クエリの作成を支援する視覚的表現を開発する。これに関連するものとして、文書収集を行うサービス、文書収集の可視化手法、可視化手法を用いた効率的な検索方法などが挙げられる。

2.1 文書の収集手法

本研究の目的における文書収集では、再現率の高い文書集合を求めている。これに関連する文書検索システムとして、論文サーベイや特許文献検索システムが挙げられる。例えば、特許リサーチでは、検索対象に関連する全ての文書を見つけることを前提とする。たった1つの文書を見残しただけでも、経済的な問題が生じ得るからだ。

論文をサーベイする時、幾つか利用されている検索サービスがある。Google は Google Scholar¹ という検索サービスを提供している。このサービスでは、検索をしたいユーザが検索要求に基づくキーワードやマイナスキーワードを構成して検索する。その他、Microsoft が Live Search Academic² という Web サービスを提供している。このサービスでは、検索したいユーザの入力したキーワードに対して、システムが関連するキーワードを提示する。しかし、この2つの検索サービスでは、マイナスキーワードをサジェストする機能がないため、文書の再現率を高めるのは困難である。

また、特許文献の検索サービスについて、USPTO 全文検索システム³ というサービスがある。USPTO はクエリセットの検索をサポートするが、クエリの作成については検索者の経験に任せるしかない。Koch らが開発した PatViz は特許の各方面のパラメータを可視化し、クエリ作成の支援や特許の分析を行うことができる [7] [8]。Giereth らは Web ベースの特許検索ツールを開発した [9]。彼らの開発したシステムでは、3つの方面から特許のリサーチを支援している。その3つとは、力指向のグラフィックアウトを利用した特許のクラスタの可視化とインタラクティブな 2D マトリックスを利用した特許の出願日と出願国の可視化とツリーマップを利用した特許の全体の可視化である。

¹<http://scholar.google.com/>, Google 社, 2006 (閲覧 2014/12/22)

²<http://academic.research.microsoft.com/>, Microsoft Research, 2008 (閲覧 2014/12/22)

³<http://patft.uspto.gov/netahtml/PTO/search-adv.htm>, United States Patent and Trademark Office, 1998 (閲覧 2014/12/22)

2.2 検索結果の提示手法

Kang らは、ブーリアンクエリの可視化、ダイナミックなクエリと検索結果のプレビューの提示、パワフルな画像検索といった機能を提供した [10]. Bast らは検索キーワードに対して関連するキーワードとそのキーワードのヒット数を提示して、検索者にキーワード入力のヒントを与えた [11]. Qvarfordt らはクエリ検索の即時プレビューを検索者に提示した. この結果、検索者がより多くの時間を検索結果の確認に利用でき、クエリ構築の効率が上がったと述べた [12].

これらの研究に共通したクエリ構築の難点は、検索者に検索対象に対して相応しいキーワードを提示することである. 検索者にとって、以前関わった検索対象を検索する時、キーワードは過去の経験から抜き出すことが可能だが、新しい検索対象を検索する時、新しいキーワードを学習しなければならない. そこで、本研究は各キーワードに具体的な実行効果を可視化して検索者に提示する. 検索者にとって、検索対象についてのキーワードをより速く獲得できる.

第3章 クエリ作成におけるインタフェイス設計の要件

本研究では、検索結果を一覧するためにクラスタリングを行い、クラスタの情報を視覚的に提示する。また、クエリセットを作成するためのキーワードの提示と、効率の良い実行を可能にする。そこで、インタフェイスの設計における可視化と操作の要件について考える。

3.1 可視化についての要件

本研究では、LDA の手法を使って、最初の検索結果をクラスタリングする。ユーザーがクラスタリングの結果を参考してクエリを構築する（図 3.1 の (a) を参照）。また、チェック時間が長くならないように、クラスタの数をデフォルトとして 10 に設定する。しかし、クラスタリングの結果によって重なりのあるクラスタが存在するため、一つのクラスタが一つのトピックを表していない。

ここで、

要件 1 クラスタリングの結果を初期状態として可視化する

要件 2 クラスタの間の関連性を可視化する

という要件があげられる。

また、LDA の手法によって、各クラスタにトピックを表すキーワードがまとめられるため、これらのキーワードがクラスタを消すためのマイナスキーワードとして使えると考えられる。しかし、マイナスキーワードの字面だけを見てクエリセットを作ることは困難である。マイナスキーワードによりどれくらいの文書が消されるかを見ながらマイナスキーワードを選択していく方が効率的と考えられる。

ここで、マイナスキーワードを提示するために、

要件 3 マイナスキーワードとして使えるキーワードを順序付きで提示する

要件 4 作成したクエリセットによって抽出された文書数を可視化する

という要件があげられる。

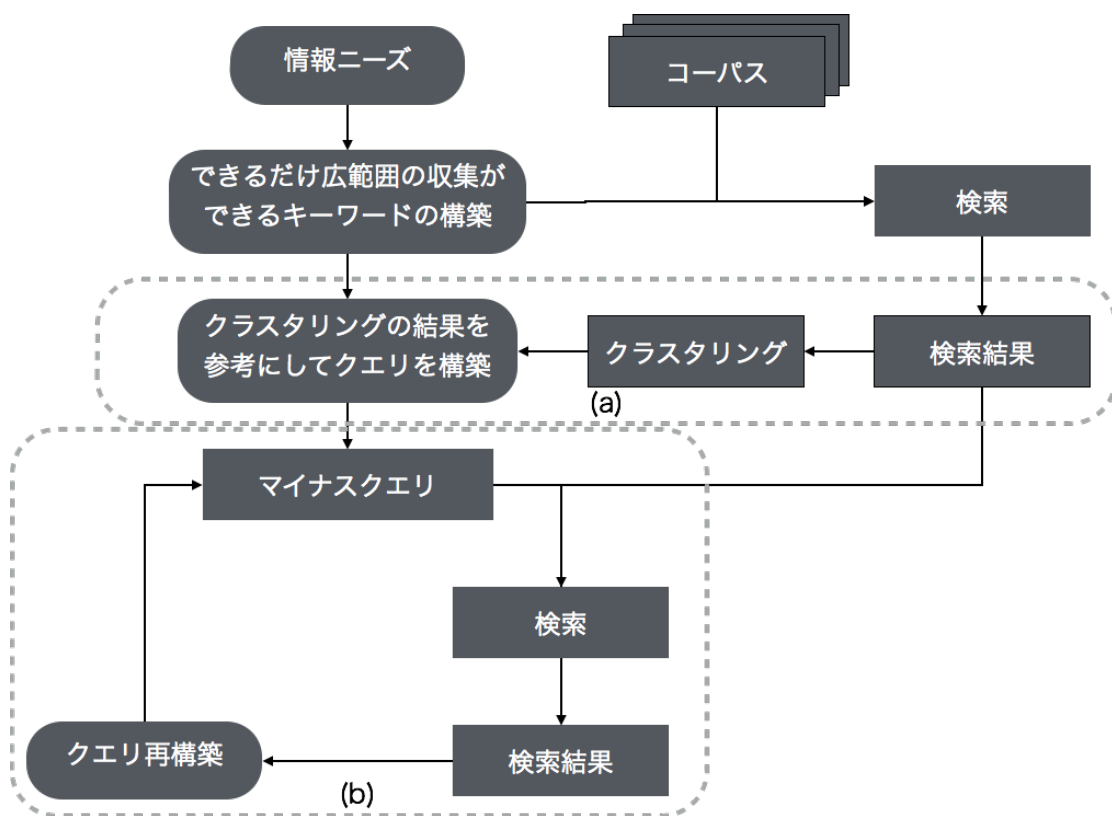


図 3.1: 本研究の情報検索モデル. (a) ユーザーがクラスタリングの結果を参考してクエリを構築する. (b) ユーザーは繰り返してクエリを再構築する.

3.2 ツール操作の要件

検索者が検索のためにクエリセットを作成する時に、クエリセットの実行効果を確認しながら、クエリセットを書き直すことが多い [1] (図 3.1 の (b) を参照)。したがって効率的にクエリセットを作るために、新しいクエリを追加あるいは削除する時、直ちに検索者にクエリの実行結果を提示することが重要と考えられる。

ここで、快適にマイナスキーワードを選択できるために、

要件 5 インタフェイス上に提示されたキーワードが瞬時に実行できる

要件 6 削除したいクラスタの選択個数は、一つでも複数でもよい

という要件があげられる。

第4章 MQViz：マイナスキーワード作成を支援するツールの設計と開発

本章では，第三章で挙げた要件に基づき，マイナスクエリ作成と支援する可視化ツールMQViz(Minus Query Visualization)の設計と開発について紹介する．

4.1 システム構成と操作の流れ

本研究はサーバとクライアントの2つのシステムで構成される（図4.1を参照）．サーバ側は検索，クラスタリングの機能が付くに対して，クライアント側は可視化とクエリ実行の機能を持つ．本研究の対象はクライアント側であるため，クライアント側の機能や操作について詳しく説明する．

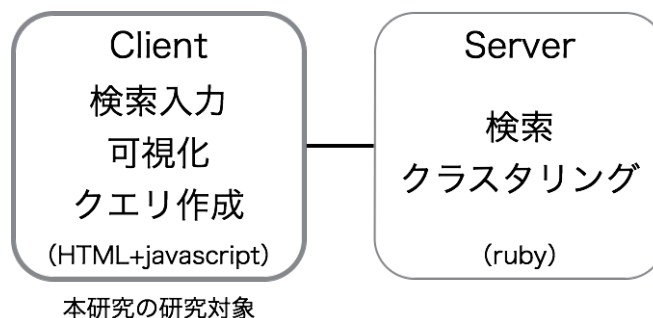


図 4.1: システム構成図

開発したシステムにおける操作の流れを図4.2に示す．まず，検索者が検索したい内容を表すキーワードを一つ入力する．サーバが検索者の入力したキーワードに基づいて文書を検索し，クラスタリングをする．そして，サーバが検索結果とクラスタリング結果をクライアント側に転送し，クライアント側がクラスタを可視化して検索者に提示する．

可視化画面では，クライアント上に観察と削除の2つのモードを実装した．観察モードはクラスタの情報を提示するモードである．このモード上で，検索者はクラスタの情報を確認することができる．そして，不要なクラスタが見つかった場合にはそのクラスタに削除のマークをつけることができる．削除モードは，必要なクラスタと不要なクラスタの関係性を閲覧することができるモードである．このモードでは不要なクラスタを効率的に削除できるキー

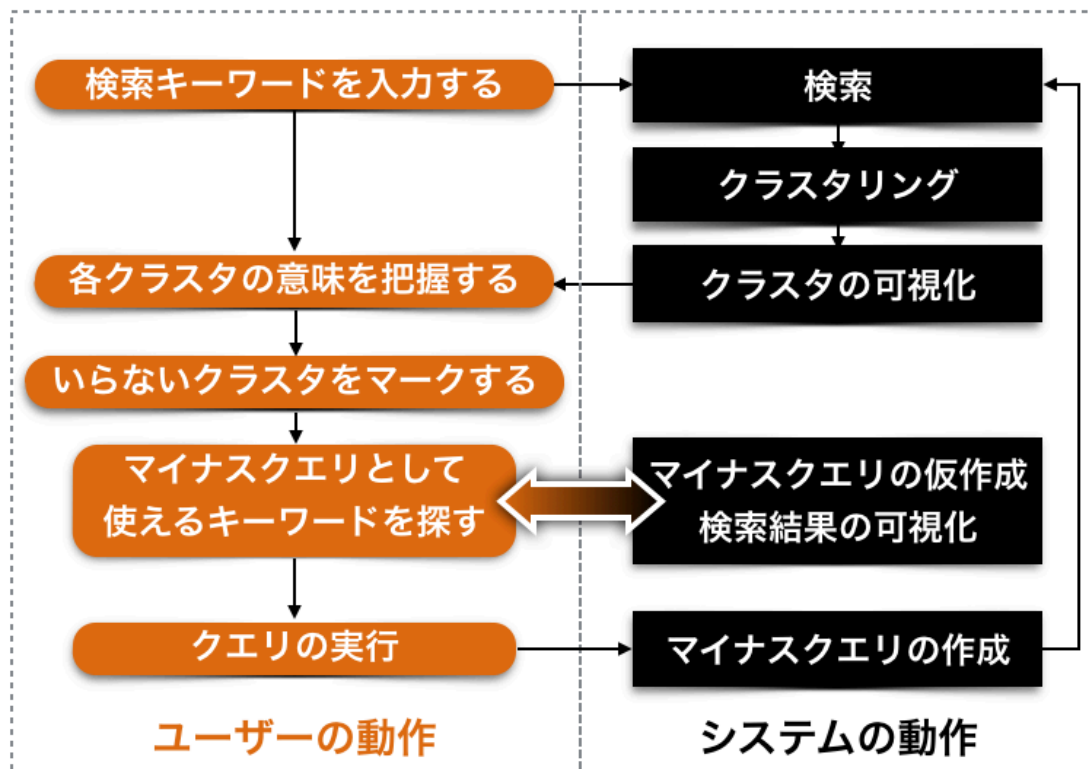


図 4.2: 操作の流れ

ワードが表示され、マイナスキーワードに指定できる。

最後に、これらの操作によって作成されたクエリセットをサーバに転送し、再検索を実行する。そして、新しい検索結果とクラスタリング結果に対して、これまでに説明した操作を反復的に行う。これによって最終的に得たいクエリセットを得ることができる。

収集された文書が分析目標に達した場合、クエリセットの作成が完成する。

4.2 ツール画面の各エリア

本ツールのインタフェースは5つのエリアから構成される（図4.3を参照する）。上側のエリアは現在の検索キーワードとユーザ操作によって作成されたクエリセットが表示されるクエリセットエリアである。左側のエリアは可視化や検索パラメータを調整するツールエリアである。可視化についての詳細は4.3節に説明する。中央のエリアは可視化エリアである。右側はクラスタの文書数、代表キーワード、スニペットなどの詳細情報を提示するクラスタ詳細エリアである。下側はマイナスキーワードを提示するキーワードアドバイスエリアである。

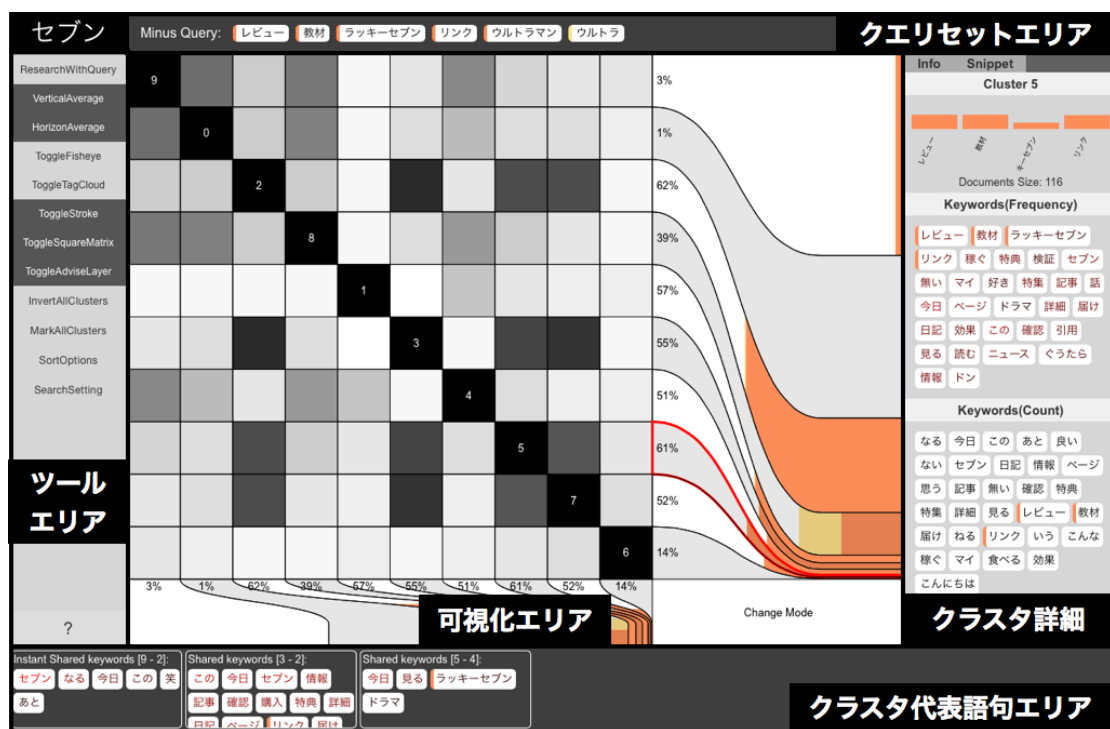


図 4.3: ツール画面上の5つのエリア

4.3 視覚的な表現

本節では、MQViz に使われた視覚的表現について説明する。

4.3.1 クラスタの視覚的表現

クラスタリングの結果を伝えるためには、クラスタの大きさ、つまりクラスタに含まれた文書数が一番重要な情報である。また、クラスタの特徴を表す他の情報（トピックスを表すキーワードや元の文章）も重要な情報である。システムでは、これらの重要な情報を視覚的に表現する。

MQViz では、クラスタを積み重なった帯状の図形で表現する（図 4.4 を参照）。長方形の可視化エリア側に隣接する辺の長さは文書数に対応している。このとき、マトリックスに隣接する辺の長さは自動的に調整される。また、文書数を表していない辺長は自由に調整できるようにする。

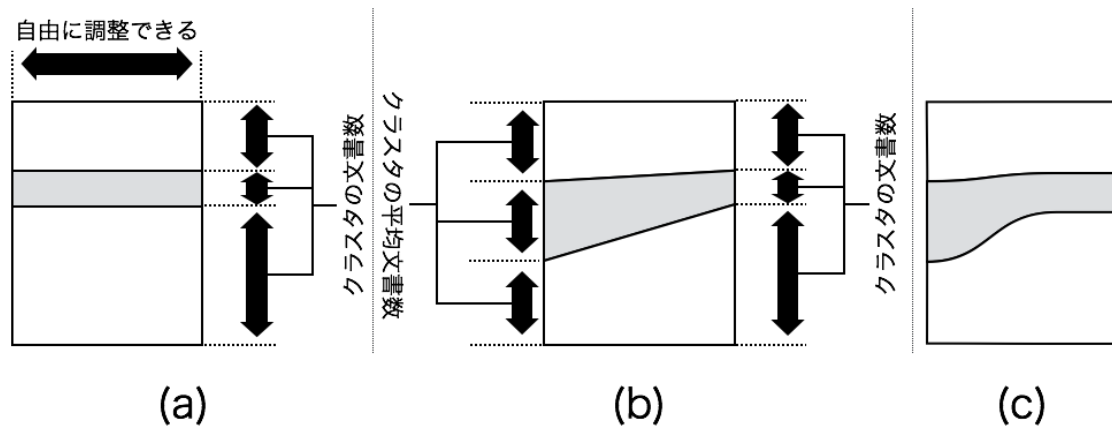


図 4.4: クラスタの視覚的表現

クラスタを水平な帯で表している時、文書数の小さなクラスタは細く表示されて見えなくなってしまう（図 4.4 の (a) を参照）。

そのため、デフォルトとして、マトリックスに隣接する辺長を平均文書数に対応させている。（図 4.4 の (b) を参照）。更に、ベジェ曲線を使って文書数を表す辺と平均文書数を表す辺を滑らかに線をつないで表現する（図 4.4 の (c) を参照）。具体的には、長方形の文書数を表していない辺長の三等分をとって、文書数に対応している一側は一つの三等分直線で描く、平均文書数に対応している 2 つの三等分線をベジェ曲線で描く。ベジェ曲線のコントロールポイントは上と下の三等分点にする（図 4.5 を参照）。

各クラスタの詳細情報として、クエリセットの実行効果、クラスタの文書数、頻度順の 30 個キーワード、カウント順の 30 個キーワードと Snippt がツールの右側の詳細エリアでまとめる（図 4.6 を参照）。

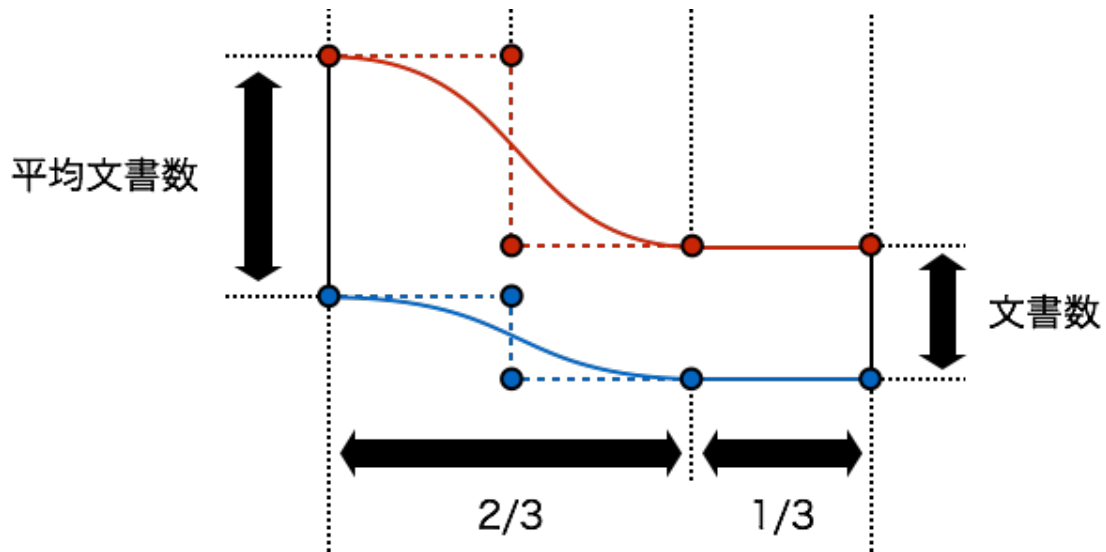


図 4.5: 左右非対称のクラスターの描画方法

4.3.2 クラスターの関連性の視覚的表現

クラスター間の関連性を各クラスターのトピックに含まれたキーワードの共有数とする。そして、可視化エリアの中にマトリックスダイアグラムを利用して全てのクラスター間の関連性を可視化する（図 4.7 を参照）。

横のクラスターと下のクラスターが交差する所の長方形が 2 つのクラスターの関連性を表す。長方形の色はキーワードの共有数に対応して、共有数が多いほど色の明度が低いのに対して、共有数が少ないほど色の明度が高い。また、同じクラスターのキーワードのすべてが共有されている場合、クラスターの関連性を表す長方形の色を黒色にする。

また、2 つのクラスターに共有されたキーワードをタグクラウドで表現する。

4.3.3 キーワードの視覚的表現

本研究が扱うキーワードは、全てマイナスキーワードに選ぶことができる。画面上に表示されたキーワードをマイナスキーワードに追加すると、自動的にクエリに追加されて実行される。

画面上に表示されるキーワードは、その目的に応じて適切な場所に表示される。あるクラスターに関わるキーワードは、画面右側にまとめて表示される。ここでは、キーワードの出現頻度順にキーワードが表示され、そのクラスターの特徴を表現している。また、同じエリアに、特定のキーワードが含まれている文書のカウント数が示される。もしある文書内に特定のキーワードが複数個含まれていた場合、出現頻度はこの個数を考慮するが、カウント数はこの個数を考慮しない。クラスターのカウント数の高いキーワードをマイナスキーワードにすれば、効



図 4.6: クラスタの詳細情報

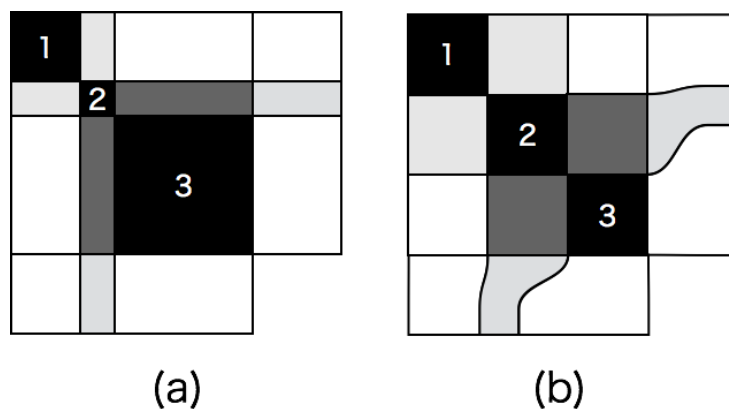


図 4.7: クラスタの関連性の視覚的表現

率的に当クラスタを消すことができると考えられる。

MQViz では、候補キーワードを4つの状態に分ける。4つの状態の遷移は図4.8に示す。

色無し 実行されていない

黄色 仮実行

オレンジ色 実行

赤色 実行効果



図 4.8: キーワードの4つの状態

仮実行は、そのキーワードがマイナスクエリとして一時的に実行されている状態である。マウスカーソルをキーワードの上においた時、キーワードを仮実行の状態にすることができる。この機能によって、ユーザにキーワードをマイナスクエリとしたときの効果を提示できるので、試行錯誤の効率を上げることができる。

実行状態は、キーワードが実際にマイナスクエリに指定されている状態である。キーワードをクリックすることによって、キーワードを実行状態にすることができる。すでに実行状態にあるキーワードをクリックした場合は、キーワードがクエリセットから削除され、キーワードが仮実行の状態になる。

実行効果の状態は、そのキーワードが各クラスタの中にどの程度含まれているかを示す状態である。この機能によって、そのキーワードがマイナスキーワードとしてどれくらいの役割を果たしているのかを確認できる。実行状態にあるキーワード上にカーソルを移動した時、可視化エリアに該当のキーワードがどれくらいの文書を消したかの情報を提示し、キーワードが実行効果の状態になる。実行効果の状態のキーワードから離れると、キーワードが実行状態に戻る。

また、キーワードを3つのエリアにわけて表示する。

- クラスタ詳細エリア
- キーワードアドバイスエリア
- 可視化エリア

クラスタ詳細エリアは、クラスタの中の頻出キーワードトップ30個を提示する。また、各キーワードの赤色の彩度によってクラスタ内のキーワードカウント数を表現する。カウント数が高いほど、キーワードの色の彩度が高くなる。カウント数が低いほど、キーワードの色の彩度が低くなる。もう一つのキーワードリストには当クラスタの中のカウント数の高い30個を提示する。

キーワードアドバイスエリアでは、観察モードでカーソルを関連性を表す長方形の上に置く時、2つのクラスタに共有されるキーワードが1つのボックスの中にまとめて表示される。そして、共有キーワードの順番は2つのクラスタの中の出現頻度の総和とする。これらのキーワードが2つのクラスタを同時に消すことができるキーワードと考えられる。

削除モードの場合、カーソルを関連性を表す長方形の上に置いた時、右側の消したいクラスタが消えて、下側の残したいクラスタが消えにくいキーワードを提示する。それらのキーワードの順番は「右側の減少文書数-下側の減少文書数」のスコアをとって、スコアの高い順にする。また、右側と下側の関連性のないクラスタのほど、キーワードが差分が大きいため、関連性のない（色が浅い）長方形から調べたほうが良いと考えられる。

可視化エリアでは、観察モードの場合(図4.9を参照)、マトリックスダイアグラムの中に2つのクラスタの関連性を表す共有キーワードをタグクラウドの形で表現する。タグの大きさが共有キーワードが2つのクラスタの中の頻度に対応する。削除モードの場合(図4.10を参照)、全てのキーワードに「右側の減少文書数-下側の減少文書数」のスコアを付き、スコアの高い30個とって、マトリックスダイアグラムの上にプロットする。また、キーワードの位置について、キーワードが右側にいるほど、右側の文書を消すことができ、下側にいるほど、下側の文書を消すことができる。

4.3.4 クエリの実行効果の視覚的表現

クエリの実行効果を瞬時に確認するために、クエリセットによる各クラスタの文書減少数を提示する(図4.11)。クエリの実行効果はクラスタを表す変形できる長方形を埋め尽くす形で表現する。埋め尽くす図形の横幅（下側のクラスタは縦幅）は文書減少の割合にする。また、クラスタの減少量を数字で提示する。

埋め尽くす図形の色について、実行されるキーワードの状態に対応して、仮実行キーワードによる文書の減少量は黄色で、全ての実行キーワードによる文書の減少量はオレンジ色で、一つの実行キーワードによる文書の減少量は赤色にする。

4.4 操作

本節で説明するクエリセット作成するための操作について説明する。

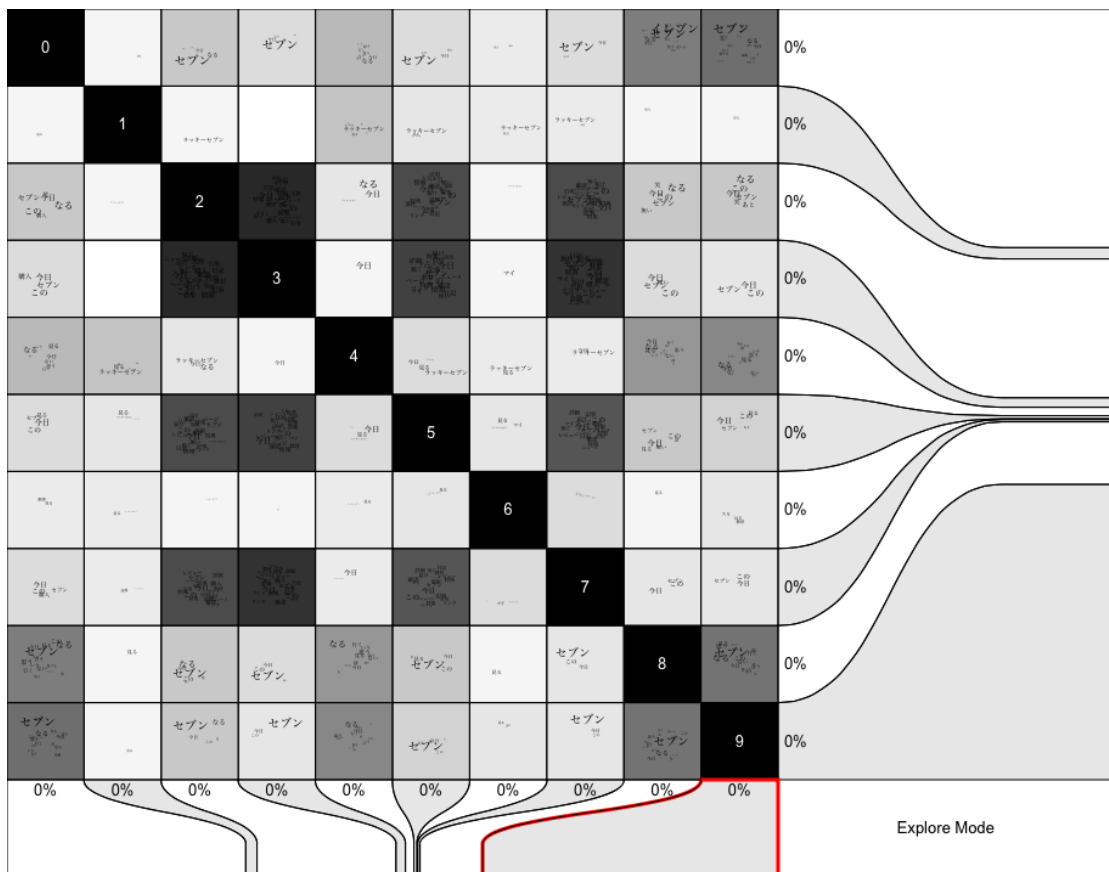


図 4.9: 観察モード

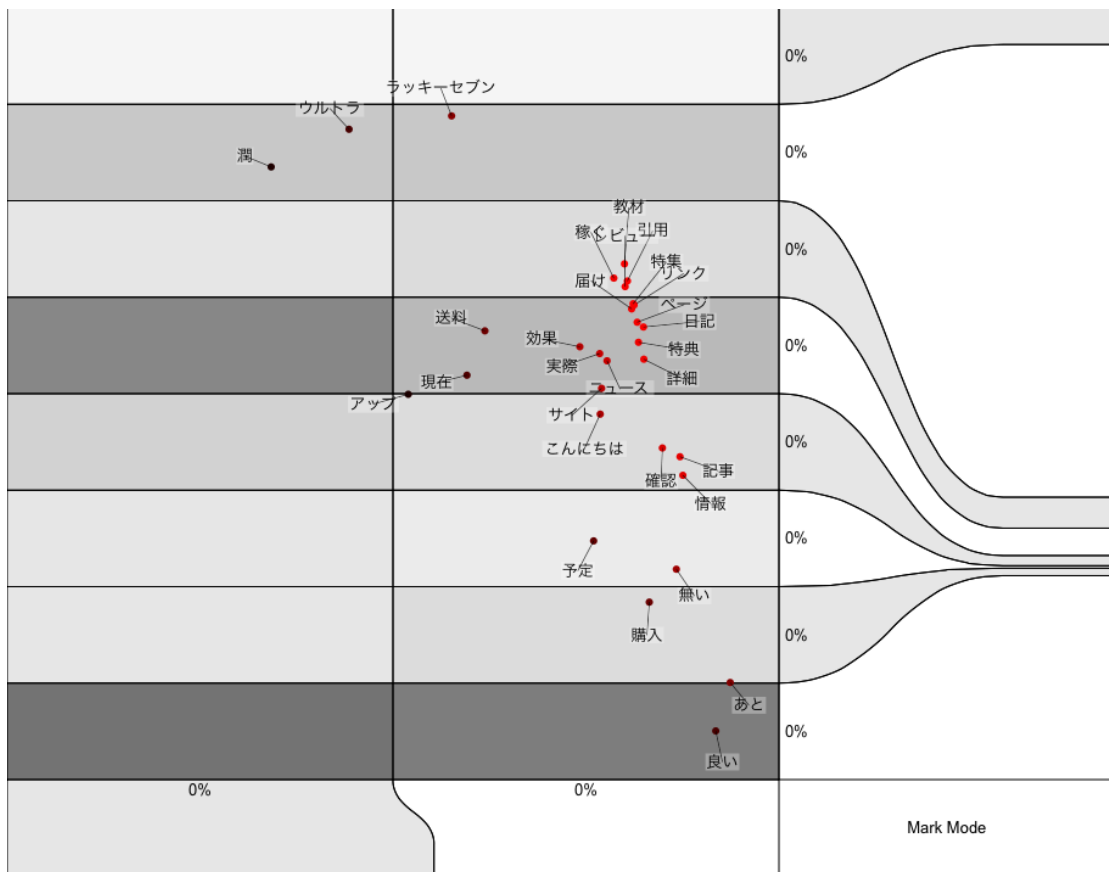


図 4.10: 削除モード

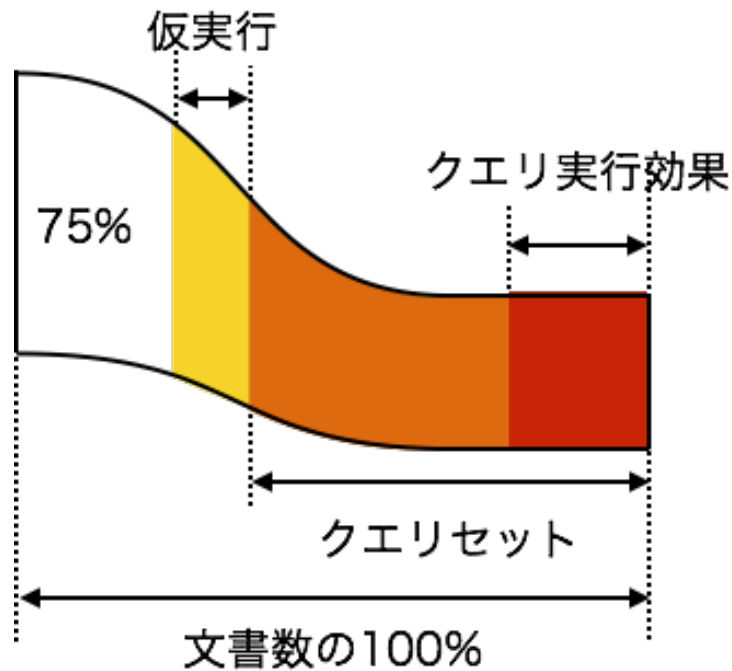


図 4.11: クエリの実行効果の視覚的表現

4.4.1 クラスタの選択

カーソルをクラスタの上に置いた時、クラスタの詳細が右側のクラスタ詳細に表示される。ユーザは詳細を見てクラスタが必要か不要かを判断する。要らないクラスタだと判断したら、クラスタをクリックする。クリックされたクラスタは不要なクラスタとしてマークされる。(図 4.12 を参照)。

また、ツールエリアの中に、「全てのクラスタを選択する」と「選択したクラスタと選択していないクラスタを交換する」ボタンを利用して、クラスタの選択を調整できる。

4.4.2 モードチェンジ

MQViz の可視化エリアは 2 つのモードに分けられる。2 つのモードの内、観察モードでは、マトリックスダイアグラムの縦軸と横軸に全てのクラスタが表示され、全てのクラスタの組み合わせについて調べられる。観察モードの状態では、不要なクラスタがマークされている時、マトリックスの右下にあるモードチェンジボタンを押すと (図 4.9 の右下を参照)、削除モードに切り替わる。

削除モードでは、先マークしたクラスタがマトリックスダイアグラムの縦軸になり、マークされていないクラスタが横軸になる。つまり、右側が消したいクラスタだけになり、下側が残したいクラスタだけになる。

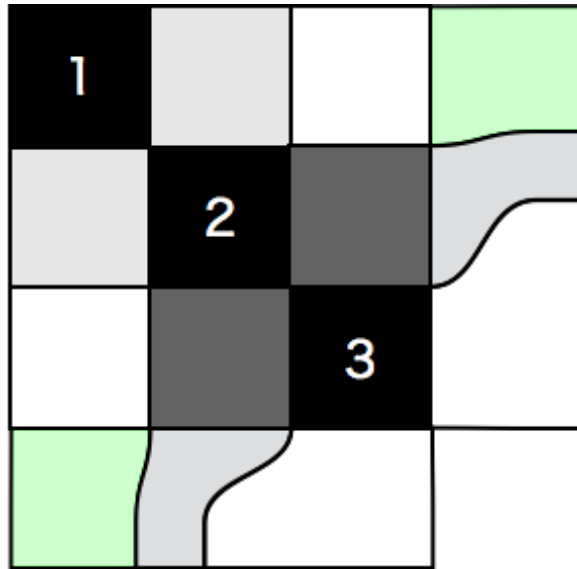


図 4.12: クラスタの選択（緑色）

また，マークされたクラスタが無い時に，モードチェンジボタンを押すと観察モードに戻る。

4.4.3 クエリの試行錯誤

クエリの効率的な試行錯誤を実現するために，単独のキーワードにホバーするだけで実行結果が確認できる他に，複数のキーワードの仮実行も対応できる．複数のキーワードをマイナスキーワードにする時，shift キーを押しながら複数のキーワードに一つ一つホバーすると，複数のキーワードがまとめて仮実行できる．また，shift キーを離すと仮実行のキーワードが全部取り消される．

4.4.4 Fisheye 機能

関連性の強さはマトリックスダイアグラムで表現する．関連性を表す共有キーワードはタグクラウドで表現する．しかしマトリックスの領域が限られているため，タグクラウドの文字サイズが小さ過ぎて見えなくなる（図 4.13 を参照）．

これを回避するため，Fisheye 機能を実装した．ツールエリアの Fisheye 機能をオンにすると，カソールの動きに合わせて，カソールが指す箇所のマトリックスダイアグラムを拡大でき，タグクラウドの文字を確認できる．

[illegible]

図 4.13: Fisheye 機能を使って共有キーワードの探索

4.4.5 補助的機能

ツールエリアにおいて、分析を補助するいくつかの機能を実装した。本節ではその機能を紹介する。

- クラスタ表現の入れ替え

4.3.1 に説明したように、クラスタの間の文書数が違いすぎる時、長方形では文書数の少ないクラスタを表すのは適切ではないため、クラスタを左右非対称の図形で表現する。それに対して、クラスタの文書数が近い時、左右対称の長方形で表現しても問題がないと考えられる。そのため、クラスタを表す2つの表現（長方形と左右対称の図形）を切り替えるボタンをツールエリアに用意した。

- 輪郭線の描画

可視化ツールは、黒い輪郭を 1px の太さの線で描画する。しかし、各ブラウザの SVG 描画エンジンが異なるため、時に、同じ 1px の直線が異なる描画結果になる。原因として、開発したシステムが非整数の座標値を持つ線をアンチエイリアスしていることが考えられる。そのため、可視化エリア内では、クラスタを表す図形とクラスタの間の関連性を表す長方形の黒い輪郭をオン、オフできる機能を追加した。

- クラスタのソート

検索した後の初期状態のクラスタ順番は、LDA のクラスタリングサーバの生成順と同じである。しかし、分析の要求によってはクラスタの表示順を調整したいことがあると思われる。そのため、ソート機能を提供する。クラスタのソート基準として、id 順 (id は LDA の結果による順番)、文書数順とクエリセットによる文書減少割合順の3つがある。文書数の大きなクラスタが検索結果として得たいクラスタ、文書数が小さなクラスタが不要なクラスタである場合が多いため、クラスタの順番を文書数順にすると、都合のよいクラスタ表示順となると思われる。クエリセットにおいて、各クラスタの影響の大きさを比較したい時には、クラスタを文書検索割合順にするとよい。

4.5 実装

本研究は全てプラットフォームを簡単に対応できるように、HTML をベースに javascript で実装を行い、Chrome、Firefox と Safari のブラウザに対応済みである。また、可視化の描画について、可視化ライブラリー d3js [13] を使用した。Scalable Vector Graphics [14] で描画を行う。

第5章 適用例

MQViz の有用性を示すため、本章は適用例を紹介する。事例として、市場調査をすることと想定して、調査員が新商品に関連する文書を収集するというタスクを設置して、MQViz を利用してクエリセットを構築する。

5.1 事例：「セブン-イレブン」についての文書収集

例えば、コンビニエンスストア大手の会社セブン-イレブンの口コミを調べるため、「セブン-イレブン」についての文書を収集することを想定する。利用するコーパスとして、ユーザの「生の声」が記録されているブログの文書が考えられる。

5.2 利用データの詳細

本事例はブログのデータを利用して、「セブン-イレブン」についてのブログ文書を収集する。しかし、「セブン-イレブン」に関する話題であっても、ブログ内の文書に「セブン-イレブン」と書くとは限らない。その代わりに、代表的なキーワードとして「セブン」で検索する。検索結果として 28631 件の文書がヒットした。次にヒットした文書を 10 個のクラスタにクラスタリングする。そして、文書データとクラスタデータを用いてクエリを構築する。

5.3 クエリセットの構築

クラスタの最初の可視化結果を図 5.1 の左側に示す。クラスタを表す帯の右側はクラスタの文書数に対応する。その中間あたりのクラスタの文書数が他のクラスタに比べて文書数が少ない。つまり、クラスタの結果がメインとなる大きクラスタと、ノイズとなる小さいクラスタだと想定される。大きいクラスタから小さいクラスタまで閲覧するために、クラスタを文書数順にソートする（図 5.1 を参照）。

そして、クラスタを 1 つ 1 つ閲覧する（図 5.2 は 2 つのクラスタの詳細を例として示す）。その結果、2 つのクラスタはコンビニエンスストアの「セブン-イレブン」に関してのクラスタ、その他のクラスタは「ウルトラマン」や「ラッキーセブン」に関して述べているクラスタである事が分かった。次に、コンビニエンスストアと関係ないクラスタをマークする（図 5.3 を参照）。

そして削除モードにチェンジする，削除モードは図 5.4 に示す．マトリックスの中のキーワードが右上にあるほど，右側のクラスタを多く消すことができ，下側のクラスタをあまり消さないことができるため，右上のキーワードから左下へとキーワードを試していく．もしくは，アドバイザエリアに提示された新しいキーワードを試していく．例えば，図 5.5 の下側は，マトリックスの中の明度の低いエリアを4つクリックした後の4つのアドバイスボックスである．その中に「安倍」「永作」「瑛」など，マトリックスエリアとクラスタ詳細エリアで提示されていないキーワードをマイナスキーワードとして試せる．

インタフェースの上の提示キーワードを参考してマイナスキーワードを繰り返し試す．図 5.5 は，作成されたクエリセットの一例として「教材，レビュー，稼ぐ，届け，日記，特典，詳細，リンク，特集，ラッキーセブン，ウルトラ，潤」をマイナスキーワードとして実行された状態である．

この後，MQViz で作成されたマイナスクエリをサーバに転送して，最初の検索キーワード「セブン」と共に検索する．そして，得られた検索結果をクラスタリングする．そこから，再び MQViz で新しいクエリセットを作成する．

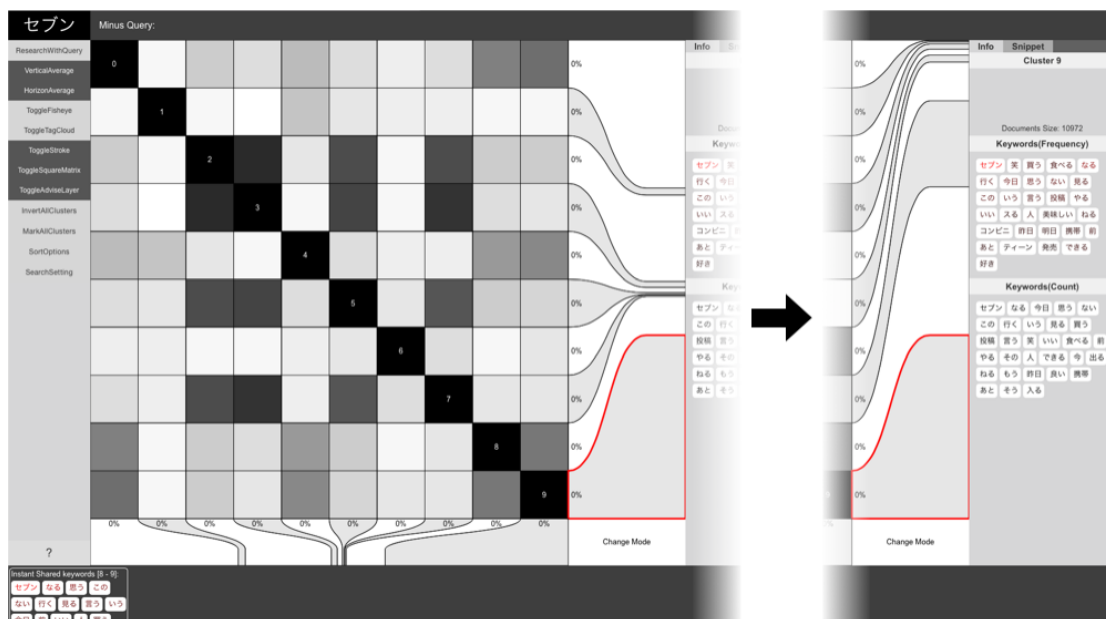


図 5.1: クラスタを文書数順にソートする

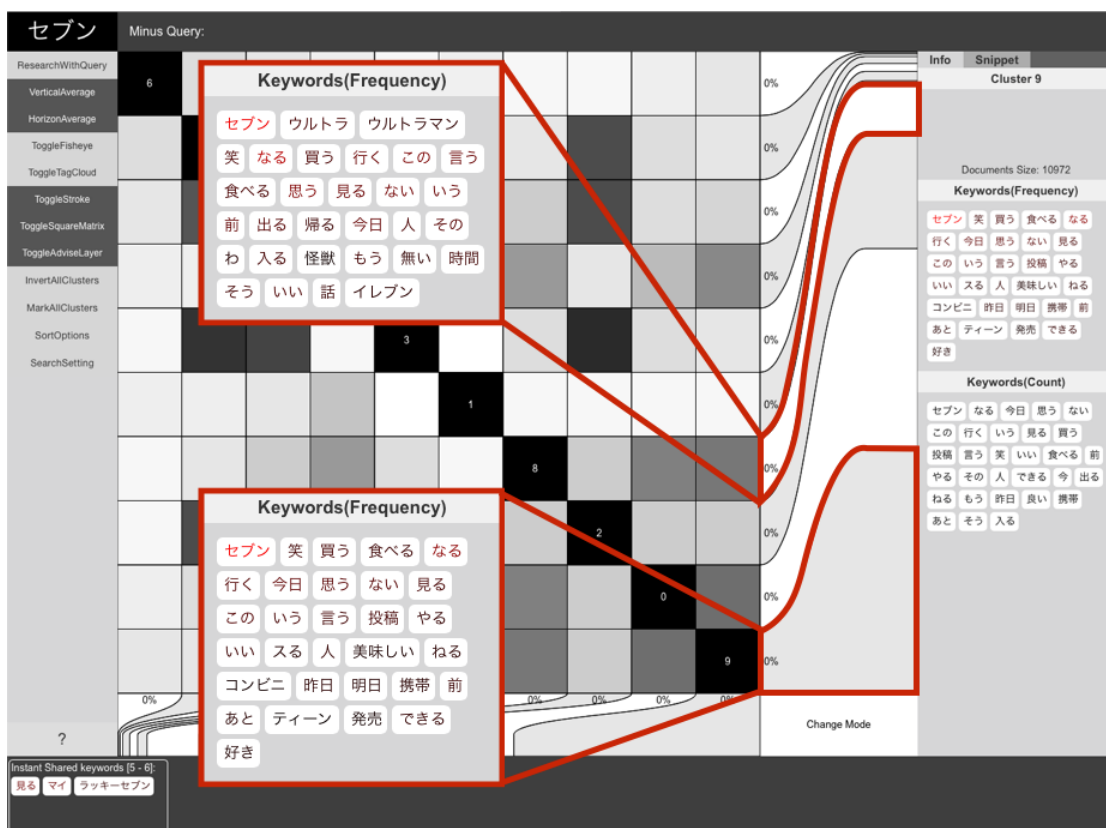


図 5.2: クラスタを1つ1つ閲覧する

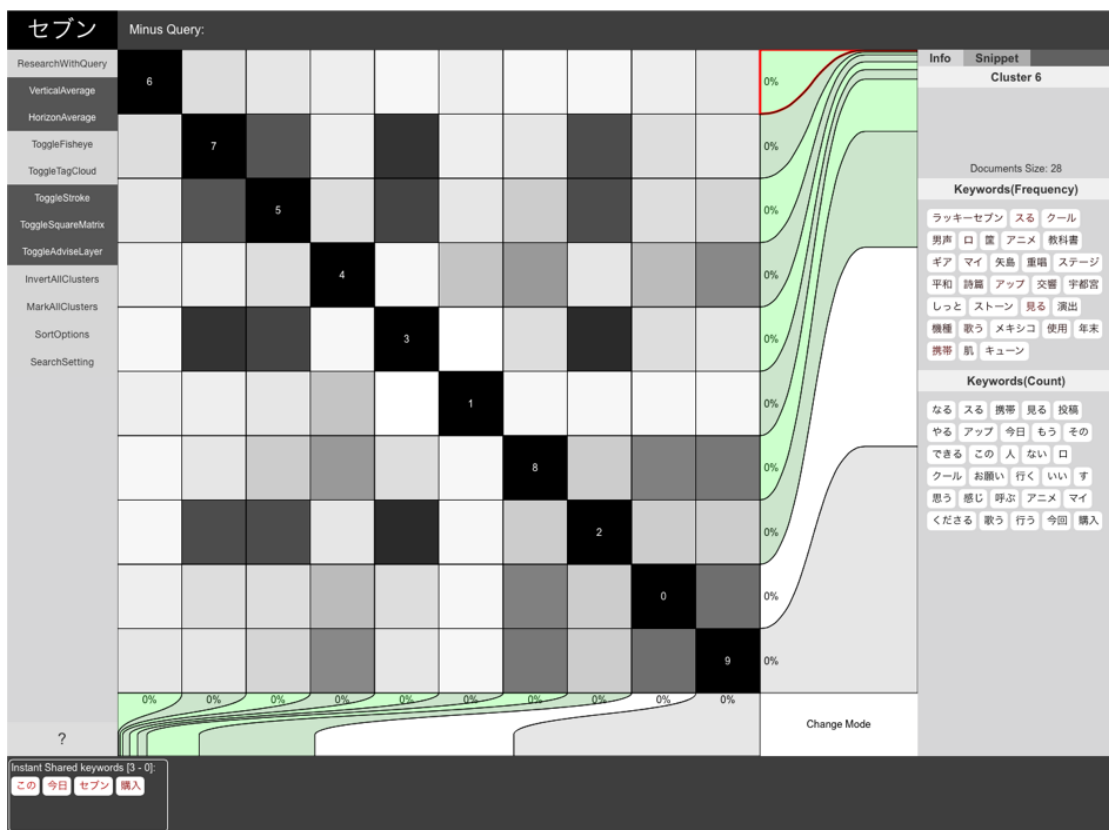


図 5.3: 緑色を重ねたクラスタはマークしたクラスタ, つまり消したいクラスタ

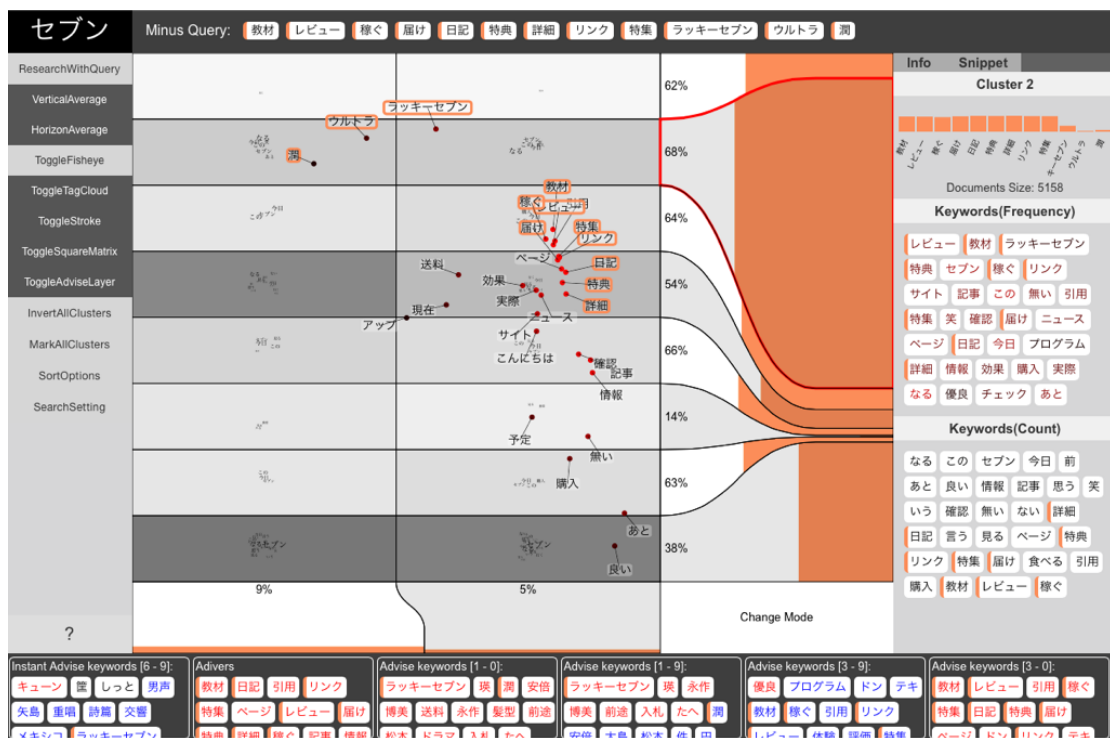


図 5.5: 削除モードでマイナスキーワードを選択する

第6章 まとめ

本研究は文書検索の目的を文書の収集と想定し、検索クエリの作成を支援する視覚的表現を設計し、可視化ツール MQViz を開発した。文書の収集方法について、本研究は1つの広いキーワードで文書を広く収集した後、マイナスキーワードを用いてノイズとなっている文書を消す検索方法を採用する。開発した MQViz は、1つの広いキーワードで収集した文書をクラスタリングした結果を元データとして使用し、複数のマイナスキーワードの提示、選択、実行を効率よく支援した。具体的には、クラスタのキーワードや、クラスタの共有キーワード、消したいクラスタを消すが残したいクラスタを消さないキーワード、などを可視化して提示することによって、重みをつけてキーワードを選択することが可能になった。また、画面上のキーワードを瞬時に実行できることによって、試行錯誤の効率を上げた。

今後の課題として、本研究が採用した「最初に一つの広いキーワードで文書を収集する」アプローチには制約があるため、その代わりに「OR クエリ」などを使って文書を収集するアプローチが考えられる。更に本研究では、クエリ作成において、マイナスキーワードの作成を主に支援したが、将来的に「OR キーワード」や「AND キーワード」などブーリアンクエリの支援も考えられる。

謝辞

本研究を進めるあたり、指導教員である三末和男准教授から、丁寧かつ熱心なご指導を賜りました。心より感謝を申し上げます。また、志築文太郎准教授、高橋伸准教授、田中二郎教授には、研究室のゼミから、様々のご助言を頂きました。深く感謝を申し上げます。

本研究のサーバとユースケースで使用したブログデータには、富士通研究所から提供いただきました。富士通研究所の皆様と様々な議論をしていただきまして、本研究がよりよく進めることができました。本当にありがとうございます。

インタラクティブプログラム研究室の皆様には、ゼミでの発表において、様々なご意見をいただきました。特に、NAIS チームの皆様には、ゼミ以外でも、研究に関しての助言をいただいております。楽しい研究生活ができまして、大変ありがとうございます。また、Tutorである小林弘明氏、後輩の武田修平氏と兵吾勇貴氏に日本語修正や日本生活のアドバイスなど、いろいろな所で大変お世話になっております。ここに、深く感謝を申し上げます。

最後に、留学を支援してくれる家族に、心より感謝しております。

参考文献

- [1] Leif Azzopardi, Wim Vanderbauwhede, and Hideo Joho. Search system requirements of patent analysts. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 775–776. ACM, 2010.
- [2] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36, pp. 3–10. ACM, 2002.
- [3] Douglass R Cutting, David R Karger, and Jan O Pedersen. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 126–134. ACM, 1993.
- [4] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 318–329. ACM, 1992.
- [5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.
- [6] Daisuke Yokomoto, Kensaku Makita, Hiroko Suzuki, Daichi Koike, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara. Lda-based topic modeling in labeling blog posts with wikipedia entries. In *Proceedings of the 14th International Conference on Web Technologies and Applications, APWeb’12*, pp. 114–124, Berlin, Heidelberg, 2012. Springer-Verlag.
- [7] Steffen Koch, Harald Bosch, Mark Giereth, and Thomas Ertl. Iterative integration of visual insights during patent search and analysis. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009.*, pp. 203–210. IEEE, 2009.
- [8] Steffen Koch and Harald Bosch. From static textual display of patents to graphical interactions. In *Current Challenges in Patent Information Retrieval*, Vol. 29 of *The Information Retrieval Series*, pp. 217–235. Springer Berlin Heidelberg, 2011.
- [9] Mark Giereth, Steffen Koch, Martin Rotard, and Thomas Ertl. Web based visual exploration of patent information. In *11th International Conference Information Visualization, 2007. IV’07*, pp. 150–155. IEEE, 2007.

- [10] Hyunmo Kang and Ben Shneiderman. Visualization methods for personal photo collections: Browsing and searching in the photofinder. In *2000 IEEE International Conference on Multimedia and Expo, 2000. ICME 2000*, Vol. 3, pp. 1539–1542. IEEE, 2000.
- [11] Holger Bast, Debapriyo Majumdar, and Ingmar Weber. Efficient interactive query expansion with complete search. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 857–860. ACM, 2007.
- [12] Pernilla Qvarfordt, Gene Golovchinsky, Tony Dunnigan, and Elena Agapie. Looking ahead: query preview in exploratory search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 243–252. ACM, 2013.
- [13] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2301–2309, 2011.
- [14] Wikipedia. Scalable vector graphics — wikipedia, the free encyclopedia, 2014. [Online; accessed 8-January-2015].