第1章　関連研究

概要

本章では、本研究に関連する研究領域について整理し、本研究の特徴や位置づけについて述べる。

1.1. 情報検索

1.1.1. 情報検索のアルゴリズム

検索アルゴリズムを提案することによって、検索システムの性能や精度を向上させようとしている研究は多い。情報検索を行うためのアルゴリズムとして、ベクトル空間モデル [?]、確率的言語モデル [?]、サポートベクターマシン [?]、Naive Bayes[] などが幅広く利用されている。

様々な情報検索手法や情報検索のアルゴリズムについては、文献 [?][?][?] などで詳しく解説されている。

1.1.2. 関連ページの検索

ある Web ページに関連したページを検索するための手法も提案されている。Chakrabarti らは、Web ページの全テキスト情報を対象として、Web ページ間の類似度を判定する手法を提案している [?]。栗原らは、ユーザのアクセス履歴を利用して、関連する Web ページ集合を検出するシステムを提案している [?]。この研究では、ユーザのアクセス履歴として、クリックしたリンクの URL とラベル文字を利用している。珍田らは、弱い紐帯を手がかりとして関連する Web ページを抽出する手法を提案している [?]。

1.1.3. ソーシャルな検索

近年では、ソーシャルな情報や人力情報などを利用した検索システムの提案も多い。本研究でもソーシャルブックマークの情報を利用した検索システムを提案している。

Agichtein らは、大規模な実験により、ユーザの振る舞いの情報を利用することによって、Web 検索の結果を改善することが可能であることを示している [?]。村田らは、検索エンジンのクリックログと検索結果として表示されるページのタイトル、スニペットを利用して、検索結果ランキングを生成するシステムを提案している [?]。White らは、ユーザのアクセス履歴を利用して、ページの重要度を決定し、人気の高いページをランキングする手法を提案している [?]。

また、ソーシャルブックマークの情報を利用した検索システムも提案されている。Xu らは，ソーシャルブックマークデータを利用したパーソナライズド検索手法を提案している [?]．さらに，ソーシャルブックマークのタグ情報を利用することによって，パーソナライズド検索を自動的に評価する手法も提案している．Yanbe らは，ソーシャルブックマークのブックマーク数を新たな指標 SBRank を提案し，PageRank と SBRank を統合して，Web 検索ランキング精度の向上を計っている [?][?]．Takahashi らは，ソーシャルブックマークデータの時間データを利用して，鮮度の高い Web ページを取得する検索手法を提案している [?]．ここでは，ブックマーク日時の散らばりの大きさから，Web ページの賞味期限を判定して，賞味期限を過ぎていない鮮度の高い Web ページを取得している．この研究は，ソーシャルブックマークの時間データに着目した情報検索手法を提案しているため，セレクトブクマの研究と近い．ただし，この研究では，鮮度の高い Web ページを取得することを目的としているのに対して，セレクトブクマでは，体系だった知識をえられ、長期に渡って役立つ Web ページを取得することを目的としている点である．そのため，時間データの利用の仕方も異なっている．

1.1.4. パーソナライズド検索

それぞれの検索者にとって最適な情報を提示するパーソナライズド検索システムの提案もされている。パーソナライズド検索では、検索者に合った情報を提示するために、よく検索者の検索履歴などの情報が利用される。パーソナライズド検索は、Google 検索 [?] にも導入されている。

Google 検索では、

1.1.5. 検索のインタフェース

新しい検索インタフェースを提案したり、インタラクティブな検索システムを提案することによって、ユーザにとってより利用しやすい検索システムを提案しようとする研究事例も多い。近年、よく導入される検索インタフェースとして、インクリメンタル検索がある。インクリメンタル検索とは、入力のたびごとに即座に検索候補を提示する検索手法である。インクリメンタル検索に関連する研究は、1990 年代から幅広くされており [?][?][?]、インクリメンタル検索の有効性も主張されている [?]。インクリメンタル検索が日本語に応用されたり [?]、音声インクリメンタル検索システムの提案 [?] もされている。

検索システムのユーザ同士のコミュニケーションを利用して、検索効率の向上を目指す研究もある。Morris らは、検索キーワードを共有したりなど、数人の小規模なユーザ間での協調作業により、情報検索の効率向上を目指している [?]。松井らは、情報と人を瞬時に発見することができ、検索サービスとソーシャルサービスの双方の利点を活用できる検索システムを提案している [?]。このシステムを利用することにより、ユーザは、検索結果と同時にページの閲覧者も発見できる。さらに、閲覧者とチャットによるコミュニケーションを行うこともできる。

渡邉らは、眺めるインタフェースを用いた検索システムを実装している [?]。このシステムでは、あらかじめ興味のあるキーワードを入力しておくことにより、入力したキーワード同士で AND 検索を行ない、検索結果が提示される。そのため、ユーザは画面を眺めているだけで、興味のある情報を得ることができる。吉田らは，検索結果ページ中に現れる重要語を話題語として抽出し，マウスによる単語のドラッグ&ドロップによって，AND 検索，OR 検索，NOT 検索を直感的に行い，検索結果の再ランキングをできる手法を提案している [?].

1.1.6. その他の検索システム

その他の情報検索システムの研究として、位置情報を利用した検索システム、ユーザの行動履歴を利用した検索システム [?]、ブログを対象とした検索システムの研究、Web 上の評判を検索するシステム、画像を対象とした検索システムの研究、連想的な検索システムの研究 [?]、2 つのキーワードの間の情報を検索するシステム [?] など様々な視点から研究が実施されている。

荘司らは、発見した Web ページを読んだ際に、読み手がどのように感じるかをクエリとして入力可能な Web 情報検索システムを提案している [?]。

1.2. 情報フィルタリング

1.2.1. 情報フィルタリングのアルゴリズム

情報フィルタリングの手法を分類すると，主にコンテンツに基づくフィルタリングと協調フィルタリングに分類できる [?].

コンテンツに基づくフィルタリングとは，コンテンツの内容に基づき情報の取捨選択を行う手法のことである．コンテンツに基づくフィルタリングに関する研究の例として，文献 [?]，[?] があげられる．

協調フィルタリングとは，ユーザの過去の行動を記録し，そのユーザと類似した行動をとっているユーザの嗜好情報から，ユーザに情報を推薦する手法のことである．協調フィルタリングに関する論文の例としては，文献 [?]，[?]，[?] があげられる．また，協調フィルタリングの実用例としてもっとも有名なのが Amazon の推薦システムである [?].

また，近年では，コンテンツに基づくフィルタリングと協調フィルタリングを組み合わせたハイブリッド法 [?] に関する研究もさかんになってきている [?].

1.2.2. ソーシャルな情報フィルタリング

ソーシャルブックマークのデータを利用した情報推薦に関する研究があり、大別すると Web ページを推薦するものとソーシャルブックマークユーザを推薦するものがある。

Web ページを推薦する研究として、Niwa らは，タグのクラスタリングをおこなうことによりタグの表記ゆれの問題の解決をはかり，ユーザのブックマーク情報からユーザの趣向に沿った Web ページの推薦をおこなう手法を提案している [?]．また、Sasaki らは，タグを表象とする Web コンテンツ群の類似性に基づいた Web コンテンツ推薦システムを提案している [?].

ソーシャルブックマークユーザを推薦する研究として、白土らは，ソーシャルブックマークユーザのブックマーク情報からユーザの関連度を解析した結果から興味の類似したユーザを推薦し，ネットワーク図として表示するシステムを構築している [?]．大力らは，ソーシャルブックマークユーザの中のイノベータ，いわばαブックマーカーに注目した情報推薦手法を提案している [?].

1.2.3. 情報フィルタリングのインタフェース

1.2.4. 様々な情報フィルタリングシステム

情報フィルタリングの研究も情報検索と同様に、様々な分野に適用されている。位置情報を利用した情報推薦の研究、画像のフィルタリングに関する研究、メールのフィルタリングに関する研究、ニュースサイトのフィルタリングに関する研究、ブログのフィルタリングに関する研究など多方面にわたる研究が実施されている。

1.3. ソーシャルブックマークデータの分析

ソーシャルブックマークや Folksonomy の分析に関する研究としては，以下のような研究事例がある．

Golder らは，ソーシャルブックマークのユーザやタグ，ブックマークの性質について分析し，各 Web ページに対する各タグの出現頻度は一定値に収束することを証明している [?]．Paul らは，del.icio.us のデータを収集して，ソーシャルブックマークが Web 検索において大きな改革を起こせるかどうか検討している．その結果，現状ではソーシャルブックマークのデータ量不足の問題やタグのゆらぎの問題から，現時点のデータでは，Web 検索に関して劇的な改革は起こせないが，今後ソーシャルブックマークのデータ量が急激に増えたりした場合は，Web 検索に革新を起こせる可能性があると結論づけている [?]．川中らは，あるタグと共起関係の強いタグを取得し，出現

時期の早いほうを親タグとする手法を用いることによって，タグの時系列の関係性をグラフ化している [?].

1.4. 検索キーワード

検索結果ではなく，検索キーワードを推薦したり、改善したりすることによって，Web からの検索を楽にしようという研究事例も多い。本研究でも長期的な検索キーワードを取得する手法を提案している。

クエリ拡張の手法として，Robertson らが提案した，検索結果から適当な単語を抽出して検索質問拡張を行う手法の一つである RSV(Robertson's Selection Value) がある [?]．正田らは，ユーザが与えたクエリでの検索結果上位 R 件を適合文書，それ以下を不適合文書として，上述の RSV を用いてクエリ拡張を行い，新たなクエリの重みにより初期の検索結果をソートする手法を提案している [?]．Zhang らは、検索エンジンのキーワード補完のように関連するクエリを推薦するために、クエリの分類と関連クエリを発見するための手法を提案している [?]。この研究では、キーワード間の TFIDF 類似度の代わりに、SF(Search Frequency) を用いた SFIDF を基づく Content Similarity の線形結合を類似度としたクラスタリング手法を提案している。Cao らは，ユーザーの入力したクエリとクリックされた URL をクラスタリングしたものと同一セッション内で入力されたクエリの関連性に基づいてクエリ推薦を行う手法を提案している [?]。大塚らは，大規模なアクセスログから抽出されたユーザの Web 検索の検索単語と，それによって閲覧したページを解析することで，関連語を抽出する手法を提案している [?]．大石らは，ユーザの意図する検索クエリを生成するための方法として，センテンス間の距離に注目した関連単語抽出アルゴリズムを提案する．このアルゴリズムは重要な語の近くに出現する単語は重要であるという考えに基づいている [?]．今井らは、入力した検索クエリが多義語の場合、選択された URL に基いて推薦クエリを生成する手法を提案している [?]。木田らは、時間とともに変化するクエリ間類似度を利用して、検索クエリをクラスタリングする手法を提案している [?]。甲谷らは、検索クエリとクリックされた URL 情報を利用して、Web サイトに到達するために頻繁に使用されるクエリを発見することで、クエリ推薦を行う手法を提案している [?]。安川らは、検索クエリのログから検索語の関連語を取得して、関連語のみに限定した単語方向のクラスタを生成し、クラスタと Web ページ群との対応をユーザに提示する Web 検索の手法を提案している [?]。

1.5. ファイルアクセスに関する研究

デスクトップ上のファイルアクセスを容易にするための研究も多くなされている。Rekimoto らは、PC 上のファイルは、すべてデスクトップに置かれ、時間とともに消えていくシステムを提案している [?]。このシステムでファイルを探す場合は、過去のデスクトップに遡っていく。そして、そのときデスクトップに置かれていたファイル群と一緒にファイルを見つけることができる。

Soules らは、複数のファイル同士の関連性を利用して、ファイル検索の精度をあげる手法を提案している [?]。

・UNIX の長期的利用ファイルへのアクセス

1.6. 時間情報の利用

時間情報を利用することによって、有益な情報を見つけようとする研究例も多い。 その中でも、流行を発見するための研究は、これまでたくさん行われてきた。

Kleinberg らは、トレンドを発見するために、時系列データの中からバーストを検出する手法を提案している [?]。Roy らは，大量のテキストデータから流行の単語を抽出する研究を行っている [?]．Ishikawa らは，時系列データを利用して，文書集合中のトレンドをとらえるための可視化システムを提案している [?].

流行現象のモデル化を行なっている研究例もある。Granovetter は、しきい値モデルを提案し、他者がどれほどの割合で流行を採用しているのか、商品を購入しているのか、という他者の採用率によって、流行現象が起こったかどうかを判定している [?]。さらに、このしきい値モデルを利用し、実際の流行現象を分析している [?]。松田は、しきい値モデルを用いて、流行がはやったり廃れたりを繰り返す流行の循環、流行が不規則にはやる流行のカオス的挙動、一度流行がはやると急激に廃れ、その後全く顧みられなくなる流行の一過性の現象などを明らかにしている [?]。中山は、流行・普及現象を、流行を採用するのか否か、流行商品を購入するのか否か、という個々人の二者択一の離散選択の集合と捉え、その離散選択をロジットモデルにより定式化を行っている。このモデルは、同調や差別化という他者の影響及び購入価格を含む採用するためのコストを考慮し、流行の採用率を算出するものである [?]。河根らは、商品などの流行現象の開始と終了を求めるために、2 次元しきい値分布を利用したモデルを提案している [?]。また、服部らは、インターネット小売業などについて、商品の流行度を反映するランキングに関する数理モデルについて解説している。そして、商品の売り上げのような社外秘に属するデータ分布をランキングという公開されたデータから分析する仕組みを提案している [?]。このように流行現象のモデル化を行ったり、流行の情報を発見したりする研究は、これまでたくさん実施されてきた。

また、情報検索や情報フィルタリングを行うために、情報が利用される時間に着目した研究例も存在する。

Dubinko らは、Flickr のタグが時間とともに発展する様を視覚化する手法を提案している [?]。視覚化手法として、川をメタファにしたものと滝をメタファにしたものを提案している。

未来の情報を予測したり、検索したりする手法も提案されている。Wolfers らは、群衆の叡智を利用して、近い未来を予測する手法を提案している [?]。河合らは、年度を明示的に含む文を検索対象とすることによって、未来情報と過去情報を検索する検索エンジンを提案している [?]。吉田らは、Web ニュースの情報を利用して、未来情報の年表を自動で構築する手法を提案している [?]。

竹井らは、情報のライフサイクルについて研究しており、文書のライフサイクルと情報のライフサイクルを対比させることにより、価値ある情報のライフサイクル管理について提案している [?]。

・ロングセラー