1D2-1

重要語抽出を用いた外部APIからの関連コンテンツ推薦

Related Contents Recommendation from Outer Contents API Using Keyphrase Extraction

近藤 光正 Mitsumasa KONDO　中辻 真 Makoto NAKATSUJI　田中 明通 Akimichi TANAKA　内山 匡 Tadasu UCHIYAMA

日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

In recent years, we can freely get various contents in contents provider from the outer contents API through inputting the search query. In this paper, we propose related contents recommendation method from outer contents API using keyphrase extraction method. An experiment shows that our method offers extracting important keyphrases as the search query in text and can recommend contents attractive to the user.

1. はじめに

近年 Web 上の情報は爆発的に増加しており，特に余暇を楽しむための動画やブログ等の CGM に関する情報増加が著しい．しかしながら余暇を楽しむためのコンテンツ探索は，ユーザは漠然とした自分の興味に基づき探索を行うため，目的指向の検索タスクと異なり検索クエリとして自分の興味を表現しにくい問題がある．そこで本研究では，与えられたテキストに対して関連した動画やブログ等のコンテンツを推薦する手法を提案する．本手法により，検索クエリを入力することなく閲覧テキストに関連するメディアが異なるコンテンツの推薦が可能になり，ユーザは閲覧テキスト以外のコンテンツから新しい気づきの情報獲得が可能になる．

従来手法の関連コンテンツ提示手法として，文書ベクトル間の類似度に基づく手法がある．この手法は与えられた文書の bag-of-words から文書ベクトルを作成し，コサイン類似度等の類似度算出手法を用いて関連コンテンツを提示する．しかしながら従来手法では，各文書ベクトルの算出と類似度計算のためにすべてのコンテンツを内部保持する必要があるため，近年積極的に公開されている外部コンテンツ API から関連コンテンツを取得することができなかった．また PLSI[T.Hofmann 99] や LSH[P.Indyk 98] 等の次元圧縮に基づき類似度算出する手法も同様の問題を抱える．そこで本研究では，与えられたテキストから重要語を自動抽出し，外部コンテンツ API が一般的に備えているキーワード検索 IF に入力することで，関連コンテンツを取得・推薦する手法を提案する．

2. 関連研究

ここではテキスト中のキーワードから関連コンテンツを提示する研究と，重要語抽出についての関連研究を述べる．Broder ら [A.Broder 07] は，Web ページに記述されている内容に関連する広告を提示する手法を提案している．彼らの手法は，各ノードに平均 100 語を含んだ計 6000 ノードから成るタクソノミーを用いて，'bid phrases' と呼ばれる広告主が指定したキーワードと一致もしくは類似する広告を提示する．その他キーワードから関連広告を提示する研究として，Wen ら [W.Yih 06] の研究がある．Mihalcea ら [R.Mihalcea 04] は，与えられた文書内におけるキーワードの共起に基づきキーワードグラフを作成し，Web 文書における情報検索ランキング手法である PageRank[S.Brin 98] に基づく重要語抽出手法を提案している．Grineva ら [M.Grineva 09] は様々な分野の文書における重要語抽出法として，Wikipedia の見出し語を重要語候補とし，与えられた文書内に含まれる重要語候補間のセマンティックグラフを作成し，密なコミュニティに含まれる重要語候補は重要であるとみなす手法を提案している．彼らの実験によると，提案手法は従来の重要語抽出法 [R.Mihalcea 04, R.Mihalcea 07] と比較して，様々な分野のテストデータにおいて精度的に優位であり，またノイズが多い Web 文書に対して頑健な抽出ができると主張している．Mihalcea らと Grineva らの手法は共に教師なし手法である．教師あり重要語抽出法として，KEA[E.Frank 99] や Hulth[A.Hulth 03] らの手法もあるが，教師あり手法の特性上，訓練データと分野の異なるテストデータが与えられた場合に精度が落ちる問題が存在する．また様々な分野の訓練データを作成するのは人手のコストが大きく現実的ではないため，様々な分野の文書において頑健な重要語抽出を行う目的の場合は，教師なし手法が理想的であると考える．本研究においては，特定の分野に特化することなく重要語を抽出し関連コンテンツを提示することを目的とするため，教師なし手法に基づく重要語抽出手法を採用する．

連絡先: kondo.mitsumasa@lab.ntt.co.jp

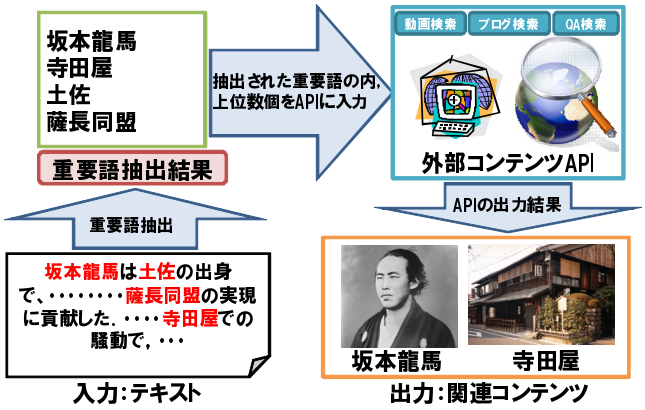

図 1: 提案手法の概要図

3. 提案手法

提案手法では，与えられたテキストから重要語を抽出し，そのうちの数語を'or' によって連結し，検索クエリ構成する．そ

して取得したいコンテンツを所有する外部 API に検索クエリを投入し，得られた検索結果の一部を関連コンテンツとして推薦する．本研究で用いる重要語抽出手法は，Wikipedia に基づく手法 [近藤 08, 近藤 10] を用いる．本重要語抽出手法は Web 閲覧履歴からの重要語抽出に用いた手法であるが，複数の文書を対象とした重要語抽出だけでなく，単独文書を対象とした重要語抽出においても精度が期待できる．次に手法の概要について簡潔に述べる．本重要語抽出手法は Grineva らの手法と同様に，テキスト中に含まれる Wikipedia の見出し語を重要語候補として抽出する．次に，共参照処理による出現頻度の修正を行った後に BM25・WebIDF に基づく出現頻度による重要度算出を行う．ここで用いる共参照処理は特に人名に特化した処理で，一般的なニュース記事では文頭で正式名称が述べられた後に，略称で呼ばれる場合が多い．(例：鳩山由起夫首相は〜鳩山首相としては〜.）すなわち略称で出現した場合も正式名称の出現頻度として数えるための処理である．そして最後に Wikipedia のリンク構造解析によって得られたキーワード固有重要度 $WKS(k)$ とポータルサイトにおける検索クエリの投入頻度に基づくキーワード固有重要度 $QKS(k)$ の線形和によるキーワード固有重要度 $KIS(k)$ を乗算することにより重要語スコアを算出し重要語を抽出する．キーワード固有重要度とはキーワードが本来もつ固有の重要度である．すなわち Wikipedia のリンク構造解析結果によって得られたキーワード固有重要度は，Wikipedia 上において Wikipedia を構築した準専門家達が考えるキーワードの重要度を表しており，検索クエリの投入頻度に基づくキーワード固有重要度は，検索された回数が多いキーワード程ユーザの興味を引きつけたキーワードであり重要であると考える手法である．以下に文書 d におけるキーワード k_d の重要語スコア $Score(d, k_d)$ の算出式を記載する．

$$Score(d, k_d) = BM25(d, k_d) \cdot WebIDF(k) \cdot KIS(k) \quad (1)$$

ここで $BM25(d, k_d)$ は文書 d におけるキーワード k の BM25[*1] の値であり，$WebIDF(k)$ はキーワード k を検索エンジンに入力することで，得られた Web 文書の総数を用いて算出した IDF 値である．ちなみに $WebIDF(k)$ と $KIS(k)$ は事前計算可能な値のため，重要語スコアの算出時にはキーワード k を key として値を取得する key/value ストアによって取得する．本研究で用いる重要語抽出手法は，Wikipedia の見出し語を重要語候補として抽出するルールベースのキーワード抽出と，出現頻度を算出後に事前計算されたスコアを格納した key/value ストアによって最終的なスコアを算出する手法のため，非常に高速な重要語抽出が可能である．そのため大規模な Web データの解析に耐えうるスケーラビリティを備えており実践的な重要語抽出手法である．

4. 評価実験

本章では，Wikipedia のキーワード候補が検索クエリとして適切であるかの評価と，提案手法の関連コンテンツ推薦に関する評価について述べる．

4.1 Wikipedia の見出し語と実際の検索クエリデータとの比較

ここでは Wikipedia の見出し語と実際の検索エンジンに投入されている検索クエリとの比較を行うことで，Wikipedia の

*1 本実験では 1 つのテキストが与えられた場合においての重要語抽出を考えているため，BM25 式における IDF 相当にあたるスコアの算出は行わない．

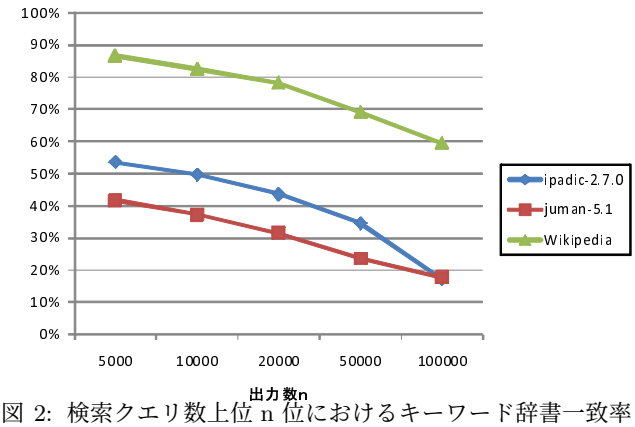

図 2: 検索クエリ数上位 n 位におけるキーワード辞書一致率（異なり数）

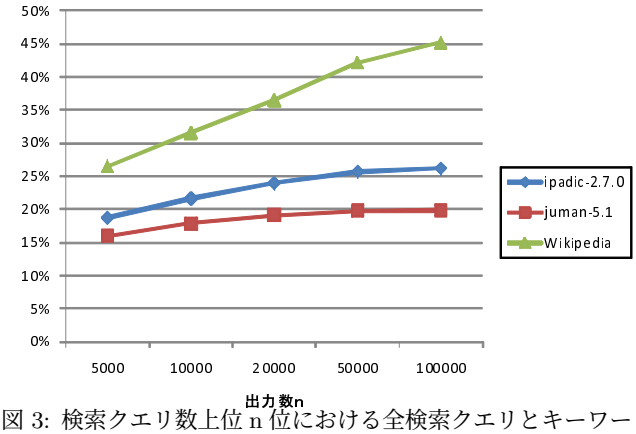

図 3: 検索クエリ数上位 n 位における全検索クエリとキーワード辞書一致率（延べ数）

見出し語が検索クエリとして適切であるかを検証する．評価に用いる Wikipedia のデータは 2010 年 1 月 7 日のデータを用い，検索クエリデータは関連コンテンツ推薦における評価実験で用いた 2009 年 7 月から 12 月末までにポータルサイト goo 上で検索されたログデータを用いた．比較対象として考えられる複合名詞や固有表現を大規模に網羅したキーワード辞書は無償では存在しないため，ベースラインとして比較するキーワード辞書として形態素解析器で用いる辞書を用いた．辞書は茶筌と Mecab で用いる ipadic-2.7.0（総語彙数約 39 万語）と JUMAN-5.1（総語彙数約 52 万語）の 2 種類を用意した．

評価方法は，検索クエリの投入頻度順に検索クエリをソートし，上位 n 位までの検索クエリに Wikipedia の見出し語と完全一致するものがどれだけ含まれるかを割合によって評価する．異なり数による評価と延べ数による評価を行い，評価式を以下のように定義した．

$$\text{検索クエリ数上位 n 位における辞書一致率（異なり数）(\%)} = \frac{\text{上位 n 位までの完全一致した辞書の異なり数}}{\text{n}} \times 100$$

$$\text{検索クエリ数上位 n 位における全検索クエリ数と辞書一致率（延べ数）(\%)} = \frac{\text{上位 n 位までの辞書と完全一致した検索クエリの延べ数}}{\text{全検索クエリ総数}} \times 100$$

本評価では，検索クエリにアルファベットと記号を含む場合，大文字小文字とバイト数を統一する正規化を行った．また複数のクエリによって構成された検索クエリは分割してそれぞれの投入数を数える評価とした．評価結果を図 2，3 に示す．

表 1: 2008 年度の重大なニュース記事を対象とした評価結果

ニュース記事タイトル	ベースライン手法				提案手法			
	重要語	動画	ブログ	QA	重要語	動画	ブログ	QA
米大統領選：米大統領、オバマ氏　４７歳・史上初の黒人、８年ぶり民主政権	0.67	0.80	**0.40**	1.00	0.67	**1.00**	0.40	0.60
クローズアップ２００８：皆既日食、離島に注目　受け入れへ懸命の準備	1.00	1.00	0.40	1.00	1.00	1.00	**0.60**	1.00
社説：リーマン破綻　危機の連鎖、米は全力で防げ	0.67	0.40	**0.80**	0.40	0.67	**1.00**	0.60	**1.00**
北京五輪：陸上　ボルト、二百も世界新　１９秒３０、２４年ぶり２冠	0.67	1.00	0.00	**0.60**	**1.00**	1.00	**0.60**	0.40
グルジア：南オセチアで砲撃、２人死亡　露が非難	**1.00**	0.20	0.20	**1.00**	0.67	**0.80**	**0.40**	0.80
北海道洞爺湖サミット：世界規模の課題が山積…　Ｇ８拡大論も	1.00	0.40	0.20	0.40	1.00	**1.00**	**0.80**	**1.00**
原油先物相場：ＮＹ　初の１４０ドル台、最高値更新　株は全世界で急落	0.67	**0.20**	**1.00**	**1.00**	0.67	0.00	0.40	0.80
地震：中国四川省震源Ｍ７．８、死者８７００人　各地で倒壊、負傷１万人超	0.67	**1.00**	0.00	0.40	0.67	0.00	**0.80**	**0.80**
ＮＨＫ紅白歌合戦：あゆで幕開け 今夜	0.67	0.20	0.00	0.80	**1.00**	**1.00**	**1.00**	**1.00**
米大リーグ：マリナーズ・イチロー、８年連続２００安打	0.33	0.20	0.00	0.40	**0.67**	**1.00**	0.00	1.00
精度の平均値	0.73	0.54	0.30	0.70	**0.80**	**0.78**	**0.56**	**0.84**

表 2: クラスタリングを用いて抽出したマイナーだと思われるニュース記事を対象とした評価結果

ニュース記事タイトル	ベースライン手法				提案手法			
	重要語	動画	ブログ	QA	重要語	動画	ブログ	QA
自動車：ＷＲＣ　ヨルダン・ラリー　ヒルボネンが今季初勝利	1.00	0.40	0.80	1.00	**1.00**	**1.00**	0.80	1.00
ことば：原爆症認定の新基準	0.67	1.00	0.20	1.00	**1.00**	1.00	**0.60**	1.00
柔らか食品：ふわふわ、もちもち人気　菓子、パン、納豆など「癒やし効果」アピール	0.33	0.0	0.0	1.00	**1.00**	**0.60**	**0.60**	1.00
世界長者番付：ゲイツ氏陥落、３位　首位は著名投資家	0.67	0.20	0.40	1.00	0.67	**1.00**	**0.60**	1.00
ネパール：排除された民、少数民族「マデシ」　インドの影響力恐れ　民主化火種に	1.00	0.60	**1.00**	1.00	**1.00**	1.00	0.80	1.00
ことば：ツォディロヒルズ	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
唐招提寺：修理完了の仏像、梱包作業始まる	1.00	**1.00**	**0.80**	1.00	1.00	0.80	0.60	1.00
登山：８０００メートル級１１座目成功の竹内洋岳さん　全１４座登頂「私の使命」	0.67	0.00	0.20	0.00	0.67	**0.60**	**0.60**	**1.00**
クローズアップ２００８：年金改ざん、底なし　「不自然な処理」公表	0.67	0.40	0.00	1.00	**1.00**	**1.00**	**0.80**	1.00
海自イージス艦・漁船衝突：石破防衛相、現場視察へ　午後、護衛艦で	0.67	0.00	0.00	0.2	**1.00**	**1.00**	**0.60**	**1.00**
精度の平均値	0.77	0.46	0.44	0.82	**0.93**	**0.86**	**0.87**	**1.00**

異なり数による評価では，上位 10000 位までの検索クエリで 80 %強が Wikipedia の見出し語と完全一致するという高い割合が示された．また，上位の検索クエリで一致しなかったものは，インターネット特有のアダルトワードや特定のサイト名が多く，表記揺らぎによる不一致も多い．しかしながら Wikipedia は表記の揺らぎを redirect 機能によって対処しているため，表記揺らぎに関してもある程度の対応が見られた．一方ベースラインとして用意した辞書は，双方とも上位 10000 位までの検索クエリで 50 %未満となる結果となった．この結果から一般的に検索されやすい固有名詞等は Wikipedia の見出し語でほぼ網羅されていることが確認された．

延べ数による評価においても，検索クエリ投入数の分布はロングテールな減衰指数関数分布であるため，上位の検索クエリほど評価割合の上昇率が高くなっている．また今回用いた Wikipedia の見出し語データは約 104 万語であるが，その内の 645683 語が評価に用いた検索クエリログ内にあることが確認された．ベースラインとして用いた形態素辞書との比較からも分かる通り，Wikipedia の見出し語は検索クエリとして用いる場合においても適切であることが確認された．

4.2 関連コンテンツ推薦に関する評価

本節では与えられたテキストから推薦された関連コンテンツの精度評価について述べる．実験では「CD-毎日新聞データ集 2008 版」から評価データを作成した．評価データとして，2008 年度の重大なニュース 10 件を Wikipedia の「2008 年」のページを参考に 10 件抽出したニュース記事データと，2008 年度の毎日新聞コーパスを記事の bag-of-words を用いてクラスタリングした結果，最もクラスターが小さくかつクラスター内の類似度が低いニュース記事を対象に各クラスター毎に 10 件抽出した計 20 件のニュース記事データをテストデータとして用いる．本評価データは重大なニュース記事からのコンテンツ推薦精度とマイナーな記事からのコンテンツ推薦精度の検証を目的としている．関連コンテンツ候補としては，動画，ブログ，QA の 3 つのコンテンツを用意した．評価実験に用いた外部コンテンツ API として，動画の取得には Youtube の Google データ API[*2]を用い，ブログと QA の取得には goo ブログ[*3]と教えて goo[*4]の API を用いた．ベースラインには，形態素名詞の連続語を重要語候補として抽出し tf によって重み付けする手法に，さらに重みが同一の場合は文頭に近い語を優先する手法（Lead 法）を用いた．そして評価方法として以下の 2 点の評価を実施した．

1. 抽出された重要語の上位 3 件のうち正解の重要語がどれだけ含まれているかの重要語の精度評価.
2. 推薦された各コンテンツ出力結果の各上位 5 件を実際に閲覧し，記事とコンテンツが関連しているかの精度評価.

ここで記事とコンテンツの関連についての極端な正解例を挙げると，ヨルダンで開催されたラリーの記事において，ヨルダンの内戦に関するコンテンツが提示された場合も，正解とした．記事としての類似性はないものの，ユーザにとってはヨルダンに関する新しい情報が得られ，記事に関してより深い洞察を得ることが出来るため，有益であると考えたためである．

∗2 http://www.youtube.com/dev
∗3 http://blog.goo.ne.jp/
∗4 http://oshiete.goo.ne.jp/

評価実験の結果を表 1，2 に掲載する．重大なニュース記事とクラスタリングによって抽出したマイナーだと思われる記事の評価実験において，提案手法が重要語抽出の精度とコンテンツ推薦の精度共に良い結果を示している．その一方で，マイナーだと思われる記事の評価実験の方が，各手法ともに重大なニュース記事を対象とした評価実験よりも良いという予想外の結果が得られた．この理由としてクラスタリングによって得られたニュース記事は文字数が少ない記事が多く，かつ記事の内容が簡潔に記述されている傾向があり，重要語候補となる語が少なく重要語抽出の精度向上に寄与したと考えられる．全体的な傾向をみると，重要語抽出で高い精度を得られた記事の関連コンテンツ推薦精度はやはり高い．しかしながら，記事内においては重要語ではあるが多義を含む語や比較的一般的な語が抽出されている場合，コンテンツ検索結果も曖昧な結果が返却される場合があるため，内容の絞り込みがしやすい重要語が抽出されている必要がある．例えばリーマン・ブラザーズの記事に関して，ベースライン手法は「リーマン」という重要語を抽出したが，このキーワードをそのまま検索した場合，サラリーマンという意味で用いられるリーマンの結果が返される．この点において提案手法は，百科事典である Wikipedia の見出し語を用いるため正式名称での抽出が可能であり，かつキーワード固有重要度によって重要な語を出現頻度に関係なく抽出できるため,「リーマン・ブラザーズ」を重要語として抽出できた．その結果，多義性を含まない正しい関連コンテンツを多数提示できている．また逆に複合名詞をすべて抽出する手法の場合，キーワードが長すぎてノイズとなってしまう場合もある．その点 Wikipedia の見出し語は長すぎず短すぎないキーワードを抽出できるため，4.1 節で述べた実験で示されたように検索クエリとして適切な語の単位を切り出すことができる．

各コンテンツの推薦精度においては，各手法とも QA，動画，ブログの順に良い結果が得られた．その理由として，QA のコンテンツは質問者が良い回答を得るために丁寧にタイトルと本文を記述している事例が多いことが挙げられる．このように検索先のコンテンツ特性による原因や，コンテンツ検索 API 側の検索精度等が精度に大きく影響している．特にブログにおいてはスパムサイト等が精度を下げる大きな原因となっている．

5. まとめ

本研究では，与えられたテキストから Wikipedia に基づく重要語抽出を行い，抽出した重要語を外部 API に入力することで関連コンテンツを取得・提示する手法を提案した．評価実験の結果，提案手法は高い精度で重要語を抽出し，また高い精度で関連コンテンツを推薦できることを示した．文書ベクトル間の類似度に基づく従来手法は，提示したいコンテンツを自社ですべて保持する必要があったが，自社ですべてのコンテンツを保持することは現実的でない．またコンテンツ空間が膨大な場合，類似度に基づく手法は計算量的にも難である．そのため本提案手法は高速な重要語抽出手法に基づく手法であり，かつ検索 API と容易に連動可能な手法のため，実践的な関連コンテンツ提示手法であると考える．今後の予定として，情報推薦の評価尺度でよく用いられるユーザにとって未知の発見的なコンテンツを推薦できているかの評価等のさらなる総合的な評価を行うことと，本評価実験で用いなかったクーポンや商品等の外部 API に対する実験や，Web 閲覧履歴からの重要語抽出手法による関連コンテンツ推薦等を検討している．

参考文献

[A.Broder 07] A.Broder, , M.Fontoura, , V.Josifovski, , and L.Riedel, : A Semantic Approach to Contextual Advertising, *In Proceedings of the 30th annual international ACM SIGIR Conference(SIGIR '07)* (2007)

[A.Hulth 03] A.Hulth, : Improved Automatic Keyword Extraction Given More Linguistic Knowledge, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '03)* (2003)

[E.Frank 99] E.Frank, , G.W.Paynter, , I.H.Witten, , C.Gutwin, , and C.G.Nevill-manning, : Domain-specific keyphrase extraction, *In Proceedings 16th International Joint Conference on Artificial Intelligence (IJCAI '09)* (1999)

[M.Grineva 09] M.Grineva, , M.Grinev, , and D.Lizorkin, : Extracting Key Terms From Noisy and Multi-theme Documents, *In Proceedings of 18th Conference on World Wide Web (WWW '09)* (2009)

[P.Indyk 98] P.Indyk, and R.Motwani, : Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, *In Proceedings of the 31th Annual ACM Symposium on Theory of Conputing(STOC '98)*, pp. 604–613 (1998)

[R.Mihalcea 04] R.Mihalcea, and P.Tarau, : Textrank:Bringing Order into Texts, *In Proceedings of 9th Conference on Empirical Methods in Natural Language Processing (EMNLP '04)* (2004)

[R.Mihalcea 07] R.Mihalcea, and A.Csomai, : Wikify!:Linking Documents to Encyclopedic Knowledge, *In Proceedings of 16th Conference on Information and Knowledge Management(CIKM '07)* (2007)

[S.Brin 98] S.Brin, and L.Page, : The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117 (1998)

[T.Hofmann 99] T.Hofmann, : Probabilistic Latent Semantic Indexing, *In Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999)

[W.Yih 06] W.Yih, , J.Goodman, , and V.R.Carvalho, : Finding Advertising Keywords on Web Pages, *In Proceedings of the 15th international Conference on World Wide Web (WWW '06)* (2006)

[近藤 08] 近藤光正, 森田哲之, 田中明通, 内山匡:HITS に基づく Wikipedia ランキングアルゴリズムとユーザ履歴を用いた個人適応型クエリ推薦, 電子情報通信学会第 19 回データ工学ワークショップ論文集 (2008)

[近藤 10] 近藤光正, 田中明通, 内山匡:MyBoom：Wikipedia に基づく Web 閲覧履歴からの興味情報推薦システム, 電子情報通信学会 ライフインテリジェンスとオフィス情報システム研究会 (LOIS) (2010)