Vol.2009-NL-194 No.7
2009/11/17

閲覧履歴におけるユーザの意図を考慮したキーワード抽出方式の提案

長野 翔一[†1]　　市川 裕介[†1]　　小林 透[†1]

ウェブ広告において，ユーザのある期間の閲覧履歴を利用して広告を配信する行動ターゲティング広告が注目されている．しかし，行動ターゲティング広告は，検索連動広告のようにユーザからそのとき調べたかったことをキーワードとして与えられることは期待できないため，ある期間の閲覧履歴から，そのユーザが探していたもの，欲しかったもの (意図) をキーワードとして抽出する技術が必要とされている．キーワードの抽出には TFIDF に代表される文書単体への重みづけを適用し，その総和を採用する従来方式が利用されるが，従来方式は，各履歴を均等に扱うため，出現する履歴が少ないキーワードは抽出は困難である，
本稿はこれらの課題を解決するため，文書分類を利用し，キーワードが出現した履歴からユーザの意図を推定する方式を提案する．また，被験者実験を通して，直前のクラスタと分析期間の履歴に共通して出現するキーワードの数が確保されていれば，提案方式が従来方式より有効であることを検証した．

Keyword Suggestion Method Considering User's Browsing Interests

Shouichi Nagano,[†1] Yusuke Ichikawa[†1]
and Toru Kobayashi [†1]

In this paper, we suggest a keyword suggestion method considering user's browsing interests in access log to overcome these problems. First, we make a hierarchical tree by using keywords appearance in access logs, for extracting lower abstraction keywords. Then 2 different abstraction levels set for contextual keywords and unique keywords, for deciding abstraction level of suggested keywords. In addition, we evaluate on effectiveness of a suggested framework by experiment results.

†1 日本電信電話株式会社　NTT 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories, NTT Corporation

1. はじめに

近年，World Wide Web におけるコンテンツが爆発的に増加しており，ウェブ広告市場が拡大している．なかでも，ユーザの一定期間の閲覧履歴を利用し，広告配信を行う行動ターゲティング広告が急成長しており，注目を集めている[⋆1]．行動ターゲティングとはクラスタリングや期間により分割された閲覧履歴から，ユーザごとに配信する広告を変化させる手法である。行動ターゲティング広告は検索連動広告のようにユーザの調べたかった対象をユーザ自身によって与えられることを期待することはできないため，一定期間の閲覧履歴から，その期間におけるユーザの意図（ユーザがその期間に探していたもの，欲しかったもの）に相当するキーワードを抽出することが求められる．

閲覧履歴内に出現するキーワードには多様な抽象度のキーワードが存在する．抽象度の高低は，オントロジーやシソーラスなどの概念構造に配置したとき，上位の概念に配置されるキーワードを抽象度が高い，下位の概念に配置されるキーワードを抽象度が低いと定義される．抽象度の高いキーワードはウェブ閲覧の特徴を捉えにくく，詳細なターゲティングを行うのは難しい．一方で，抽象度の低いキーワードは特徴的なキーワードが多いが，必ずしも閲覧の意図とは合致していない．そのため，行動ターゲティングに利用するためには，適切な抽象度を決定するキーワード抽出方式が必要とされる．

文書群からキーワードを抽出する提案は数多く行われている．しかし，既存手法の多くは出現履歴数が多いほど，優先して抽出される傾向にあるため，出現履歴数の少ない語を抽出することは困難である．抽象度の低い語の多くは出現履歴数も少なく，概念構造を用いなければ抽象度の低いキーワードは抽出されないことが多い．

一般的に，抽象度の決定には，オントロジーやシソーラスなどの概念構造が利用される．しかし，新語への対応や，分野ごと，ユーザごとのカスタマイズはコストが大きいという問題がある．

そこで，我々は抽象度を算出することなく，同等の効果を得るキーワード抽出を提案するため，既存方式で得られるキーワードの出現傾向から，出現する履歴数が多いほどキーワードは抽象度が高く，出現パターンが類似したキーワードは概念的に近いという仮説を設定し，キーワードが出現する履歴の類似性から概念構造を模したツリーを作成する．

また，抽出すべきキーワードの抽象度の決定も課題となる．そこで，我々はユーザの行動が閲覧期間によって探している対象の抽象度も変化していることに注目した．たとえば，テレビの購入を検討している一連の閲覧履歴において，ユーザの意図も「テレビ」「プラズマテレビ」「テレビ」「ハイビジョンテレビ」「ブラビア」というように直前の閲覧期間の意図を踏まえて対象の抽象度が上下に変化していた．このウェブ閲覧の性質を利用すれば，直前の閲覧期間のキーワードを参照し，その差分を利用することで，閲覧の文脈を考慮した抽象度を決定することができると考えた．

本稿の構成について以下に説明する．

⋆1 「インターネット検索エンジンの現状と市場規模等」に関する調査結果 (2009 年総務省) によると 2013 年には 2005 年の 8.7 倍にあたる 841 億円の市場規模になると予測されている

Vol.2009-NL-194 No.7
2009/11/16

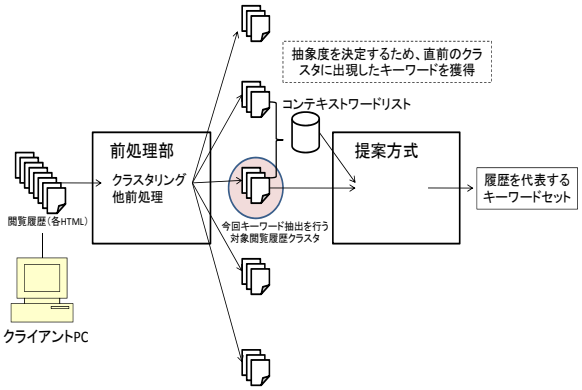

図 1 クラスタリングと連携するキーワード抽出のフレームワーク

はじめに，2 章において，関連技術の紹介について示す．3 章において，提案方式のアルゴリズムについて説明し，クラスタリング技術と組み合わせたシステム構成の一例を示す．4 章において，提案方式のフレームワークが有効に機能することを検証した評価実験について示す．5 章において，評価実験における議論について示す．最後に 6 章において，本稿のまとめについて示す．

2. 背　　景

2.1 取り組む課題

本稿では，抽象度を考慮したキーワード抽出を実現するために，2 点の課題の解決を試みる．

課題 1　概念構造を用いず抽象度を図る基準を設定する必要がある
課題 2　抽出すべきキーワードの抽象度の決定する必要がある

2.2 関 連 研 究

複数の文書から重要なキーワードを抽出するために最もよく使われているアプローチは，TF，IDF[1)]，LR[2)]，c-value[3)4)]，BM25[5)] といった文書単体へのキーワードの重みづけ手法の総和をとり，拡大適用する方式である．これらの手法は文書に出現する語句に重みを与えることで対象文書から重要なキーワードを抽出することを可能とした．しかしながら，複数の文書で構成される閲覧履歴へ適用すると先にあげた問題が発生する．また，適切な文書群を基にした IDF 値を適用することで，少ない文書で出現するキーワードを抽出されることがあるが，抽出するキーワードの抽象度を考慮していないため，抽象度を決定するためには概念構造を利用する必要がある．

ユーザの意図と合致するキーワードを抽出可能とするため，文書群からユーザに興味のある文書や，興味のあるキーワードを選択してもらう方式[6)7)] も研究が行われている．しかし，行動ターゲティング広告への適用を想定すると，ユーザとのインタラクティブな入力は期待できない．

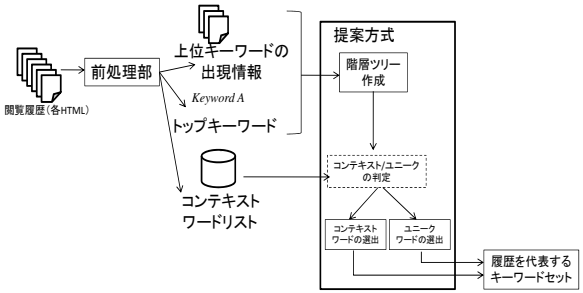

図 2 処理の流れ

キーグラフ[8)]，CF 法[9)] は，従来の重み付け方式がユーザの意図を考慮していないことを課題としており，頻出語との共起関係により重要度を付与することで，特徴的なキーワードの抽出を可能とした．これらの研究により，文書の頻出語と各語の共起関係が重要度を求める指標として有効であることが示された．このように，キーワード抽出において，共起関係をはじめとしたキーワードの出現，非出現情報の有効性は広く知られており，本研究においても，各履歴における語の出現情報からツリー作成を行い，ユーザの閲覧意図と関係の薄いものを除去することを試みる．しかし，これらの方式はしばしば，一般的な語 (抽象度の高すぎるキーワード) を抽出することを自身の論文で指摘しており[10)]，本稿における課題 2 を解決するためには，シソーラスなどを用いて概念レベルでの処理を行う必要がある．

chen08[11)] はウェブディレクトリから Concept hierarchy という階層的な概念構造を作成し，広告マッチングに適切なキーワードを抽出する方式を提案し，その有効性を示した．彼らは Concept hierarchy の作成にウェブディレクトリを利用しており，抽出キーワードを広告マッチングに特化させている．このように，概念的な上位，下位を考慮したキーワード抽出を行うためには，階層構造のツリーを作成するアプローチが有効であると考えられる．しかし，彼らの概念構造の構築手法はウェブディレクトリの構造に依存しており，対象とするウェブページの構造が抽象度に基づいて構築されていない場合，適用が困難となる

以上のように，従来のキーワード抽出において，キーワードの抽象度に着目し，人手や与えられた概念構造を介することなく，抽象度を自動決定する研究は行われておらず，本研究における新規性となる．

抽出すべきキーワードの抽象度については，特にオントロジーやセマンティックウェブの分野で抽象度の不一致の問題として研究がおこなわれている．Kuhlthau91 が提案する ISP モデル[12)] は情報探索過程の進行を表現しており，レボウィッツ 07[13)] は同モデルに基づき，提示情報の抽象度を段階的に下げる方式を提案している．このように，抽象度決定に情報探索の遷移が重要であり，我々は，複数の情報探索過程が並存する可能性があり，情報探索過程の逆行（焦点形成から探求へ逆行するなど）などが考えられるウェブの閲覧履歴におい

Vol.2009-NL-194 No.7
2009/11/16

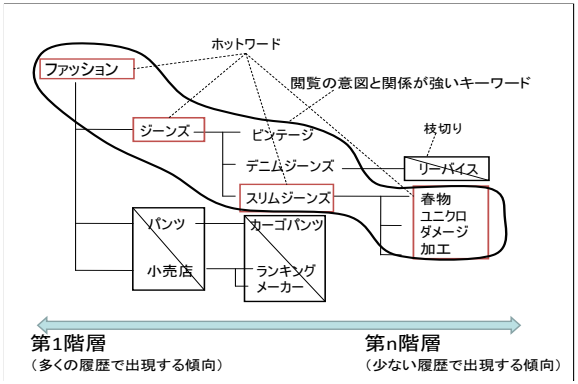

図 3 階層ツリーサンプル

て，情報探索過程の遷移を抽出するため，隣接する二つの期間の閲覧履歴の差分を利用する方式を提案する．

3. 閲覧履歴を要約する代表語抽出の提案とその実現

提案方式の処理は，ツリー作成とキーワード選出の二つのステップに大別される．初めに，キーワードの出現情報からキーワードの階層ツリーを作成することを目的とし，ユーザの意図と関係の深いキーワードをホットワードとして定義し，それらを抽象度ごとに連ねた階層構造を階層ツリーと呼ぶ．階層ツリーは，図 2 のように，高い階層から順に第一階層，第二階層... 第 n 階層で構成され，それぞれの階層にキーワードを含む構成となる．次に階層ツリーを構成するキーワードをコンテキストワード，ユニークワードに色付けし，ホットワードを中心に異なる方法でそれぞれ指定された数のキーワードを抽出する．

提案方式はキーワードの出現情報，トップキーワード，コンテキストワードリストを入力とし，指定数のキーワードを出力する．また，コンテキストワードを獲得するため，クラスタリング技術と組み合わせて使用される．すなわち，提案方式は一つのトップキーワードで表現される適切な履歴数で構成される (5〜25 程度) 閲覧履歴を処理対象とし，直前に見ていた履歴群のキーワードを獲得できることを必要とする．これらの前提は，クラスタリング技術との連携により，達成される．

3.1 キーワードの出現情報を利用したツリー作成

ツリー作成は各キーワードがどの履歴で出現しているかを示す出現情報と入力された閲覧履歴全体を表現するトップキーワードを入力とし，キーワードで構成された階層ツリーを出力する．作成される階層ツリーは浅い階層ほど出現履歴数が多く，階層が深くなるほど出現履歴が少なくなる傾向にあり，また，階層が隣接し，親子関係を有するキーワードは出現パターンが類似する性質を持つ．ツリー作成の仮説に基づくと，この階層ツリーは，抽象度

Algorithm1-MakeTree
Input: a new value $TopWord$, $EI(log - num, keyword)$ as Emerging information
Parameter Setup: $N(0 \leq N \leq 1)$
Output: @hotlist and @keylist as Concept Tree

```
1.  Hotword = TopWord;
2.  @hotlist ← Hotword
3.  @keylist ← Hotword
4.  Process = continue;
5.  unprocess = all of EI keyword;
6.  while process = continue do
7.    @list = kwd;
8.      matching N ≤ Score(Hotword, kwd) from unprocess
9.    remove @list from unprocess
10.     if @list=0 do
11.       process=END;
12.   Hotword ← max (number of word);
13.     number of word = @ (N ≤ Score (@list, unprocess))
14.  @hotlist ← Hotword
15.  @keylist ← word
16.    word matching from @list(higher than Hotword's Score)
17.  Report @hotlist and @keylist;
```

Algorithm2-KeywordExtract
Input: @hotlist and @keylist as Concept Tree, @context as Context word list ,c-keynum, u-keynum
Output: @output

```
18.  foreach kwd (@hotlist)
19.    if kwd ∈ @context do
20.      @clist ← kwd;
21.    else do
22.      @ulist ← kwd;
23.  @output ← c-keynum of word
24.    extracted from @clist top
25.  @output ← c-keynum of word
26.    extracted from @clist bottom
27.  Report @output
```

図 4 Algorithm

の高いものを浅い階層に，抽象度の低いものを低い階層に配置し，概念的に近いキーワード同士を隣接する階層に配置した擬似的な概念構造である．ツリー作成のステップは以下の手順で行われる．

手順 1 最も浅い階層（第 1 階層）にトップキーワードを配置する．トップキーワードは TFIDF 値の総和などにより決定される．

手順 2 第 1 階層のキーワードを親キーワードとし，まだ配置されていないキーワード (子キーワード候補) との出現情報の類似性を式 1 に定義されるランキングアルゴリズムにより策定し，閾値 N を超えたスコアを有する子キーワードを第 2 階層に配置する．

Vol.2009-NL-194 No.7
2009/11/16

手順 3 第 2 階層の各キーワードを親キーワードとし，まだ配置されていないキーワード (子キーワード候補) との出現情報の類似性を式 1 に定義されるランキングアルゴリズムにより策定し，閾値 N を超えたスコアを有する子キーワードを親キーワード直下の第 3 階層に配置する．このとき，第 3 階層に配置されたキーワード数が最も多い第 2 階層キーワードをホットワードとし，第 3 階層に配置されたキーワードは親キーワードがホットワード，親キーワード自身のランキングスコアを優先しながら，重複を削除する．

手順 3 手順 2 を子キーワードがなくなるまで，階層を深めて繰り返し，階層ツリーを作成する．

手順 4 階層ツリーからホットワードよりランキングスコアの低い同階層のキーワードとホットワードと異なる親を持つ同じ階層のキーワードを削除する．

親キーワード W_1 と子キーワード W_2 の出現情報の類似性を図るランキングアルゴリズム $Score$ は以下のように設定する．

$$Score(W_1, W_2) = P(W_2|W_1)(1 - P(\bar{W_2}|\bar{W_1})) \quad (1)$$

処理は出現履歴数が多いキーワードから，出現情報が類似したキーワードを集め，その中でユーザの意図となるホットキーワードを探す処理を繰り返すことで行われる．本処理において，トップキーワードという出現履歴数が多いキーワードを与えることで，処理を繰り返すほど，より出現履歴数の少ないキーワードが配置されることを期待する．

処理において子の数が多いものを残すのは，ユーザの意図となるキーワードほど多様なキーワードで表現されるため，概念的に近いものが多くなると考えたためである．

出力された階層ツリーはトップキーワードを最も浅い階層とし，各階層においてホットワードとホットワードよりスコアの高かったキーワードが存在する (第 1 階層のトップキーワードと最も深い階層のキーワード群は全てホットワードとする)．

3.2 ユニーク・コンテキストベースのキーワード選出

キーワード選出はツリー作成のステップで生成された階層ツリーとコンテキストワードリストを入力とし，所定のキーワード抽出数のキーワードを出力する．コンテキストワードリストとは処理対象の閲覧履歴より以前からユーザが興味を持っていたキーワード群であり，たとえば，1 日分の履歴を 1 クラスタとする場合，前日の履歴に含まれているキーワード群，クラスタリング技術と組み合わせて使用している場合，処理中の履歴の直前の x 個の閲覧履歴を含むクラスタがに含まれるキーワード群が利用される．すなわち，コンテキストワードとは処理対象となる履歴とその直前の履歴の共通分にあたり，処理対象の閲覧履歴に存在するキーワードの中からコンテキストワードに含まれないキーワード (ユニークワード) が差分にあたる．キーワード選出のステップについて以下に説明する．はじめに，階層ツリーに含まれるキーワードをコンテキストワードリストに含まれるコンテキストワードと含まれないユニークワードに分類する．次に，コンテキストワードは深い階層から順に，ユニークワードは浅い階層から順にキーワードを選出する．抽出の際の優先度は，ホットワード > 階層の浅深 > ランキングスコアの順で処理される．

ツリー作成時点でユーザの意図と関係の薄い枝を排除することで，直観に反したキーワードの抽出を抑止し，残ったキーワードをコンテキストワードとユニークワードに分け，コンテキストワードの深い階層にあるキーワードとユニークワードの浅い階層にあるキーワードを抽出することで直前の閲覧状況を考慮した抽象度でキーワードを抽出する．

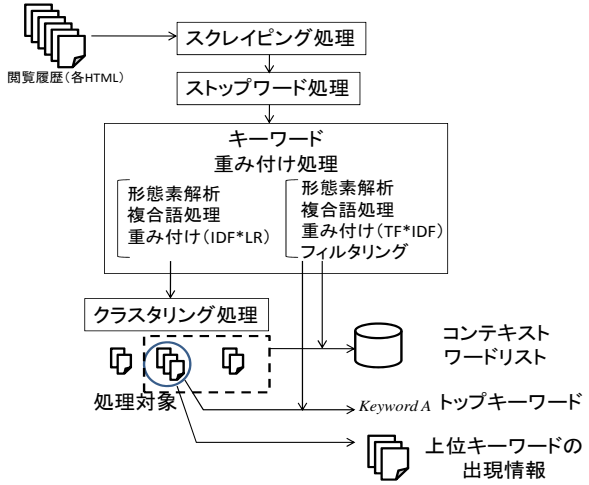

図 5 前処理部の詳細

3.3 提案方式を実現するシステム構成の一例

提案方式をクラスタリング技術と連動させるシステムの構成について示す．提案方式はクラスタリングとの連携を前提としており．以下に示す前処理を必要とする．

提案方式を運用するシステムはクライアント PC に蓄積された時系列に並んだウェブページの HTML ファイルを処理サーバに送り処理を行うことを想定している．代替構成として，プロキシサーバやリクエストサーバに履歴蓄積装置を置く構成，クライアント PC に URL リストを保存し，処理サーバで URL リストをもとに HTML を獲得する構成などがあるが，閲覧ウェブページに制限がかからない，パーソナライズ機能を有するウェブページも閲覧時の HTML が変化しないといった点から，リアルタイム性が求められない今回はクライアント PC に HTML ファイルを蓄積する構成を採用した．

3.3.1 スクレイピング処理

HTML ファイルから広告，メニューバー，フレームに当たる部分を除去し，<> で括られた HTML タグを除去する．ユーザが閲覧していた本文部分を出力する．タグ除去とスクレイピングの一部には Extract Content[14] を利用した．また，スクレイピング精度向上のためアクセスの多いいくつかのサイトについては，ルールベースでスクレイピングを行う．

3.3.2 ストップワード処理

ウェブにおいて頻繁に使われるキーワード（トップページ，規約，ヘルプ，コンテンツなど）を除去する

Vol.2009-NL-194 No.7
2009/11/16

3.3.3 キーワード重み付け処理

本処理では，1 閲覧履歴をキーワードの出現情報により表現するため，形態素解析，複合語抽出，重要度の重み付け，フィルタリングを行う．

はじめに，提案方式で対象とする名詞（単名詞，複合名詞）を抽出するため，テキストに形態素解析を行い，名詞を抽出したのち，品詞情報 (名詞が連続している) からルールベースで複合語抽出を行う．形態素解析には MeCab[15] を利用し，追加辞書に wikipedia のタイトルとなる名詞を登録した．また，複合名詞抽出には TermExtract[16] を利用した．

次に，抽出されたそれぞれの単名詞，複合名詞に対し，重みを付与する．キーワードの重みはトップキーワードの獲得，上位キーワードの出現情報の獲得，クラスタリング処理に利用されるが，トップキーワードの獲得，上位キーワードの出現情報の獲得には TFIDF を，クラスタリングの際は IDF とバイグラム連接コストの積をそれぞれ採用する．重要度の重みづけ方式は試行錯誤により，キーワード抽出の結果が適当なものを選択した．なお，IDF，バイグラムコストの算出には入力された全ての閲覧履歴を対象に学習を行った．

最後に，提案方式，キーワード重みづけの処理における計算コストの低減，問題の簡略化を考慮し，クラスタリング処理以外へ送る各履歴については，重み上位 10 キーワード以外を除去することでフィルタリングを行った．

3.3.4 クラスタリング処理

ベクトル空間法に基づき，各履歴間のコサイン類似度を算出し，クラスタリングを行う．閲覧履歴のクラスタリングは，過去の実験の結果から，クラスタ内のデータ数が 5〜25 個程度で構成されることが想定されるため，1 クラスタを構成するデータが少ない時，有効な分類を行うことのできるクラスタリング方式[17] を採用した

本処理により作成される各クラスタに対し，提案方式によるキーワード抽出が行われる．

クラスタリング処理により獲得したクラスタ内の各閲覧履歴における TFIDF 上位 10 キーワードの出現情報を対応し，上位キーワードの出現情報を獲得する．

処理中の履歴の直前の閲覧履歴を含むクラスタがに含まれるキーワード群をコンテキストワードリストとして獲得する．

いずれかの履歴上位 10 キーワードに存在しているキーワードのなかから，TFIDF 総和が最も高いものをトップキーワードとして獲得する．

以上の前処理により獲得した上位キーワードの出現情報，コンテキストワードリスト，トップキーワードを提案方式の入力とする．

4. 評 価 実 験

4.1 評 価 方 法

提案方式はユーザの意図を考慮したキーワード抽出を行うことを目的としている．提案方式のフレームワークの妥当性を評価するため，評価実験を行った．キーワード抽出方式の評価として Prescision，Recall など，抽出されるキーワードの性質に関する評価方法なども存在するが，今回の実験では，提案方式を組み合わせたフレームワークの妥当性評価を目的とするため，被験者自身により各手法で抽出されたキーワードセットをを 5 段階のリッカート尺度[⋆1]により評価することで，キーワードセットの総合的な評価を行った．

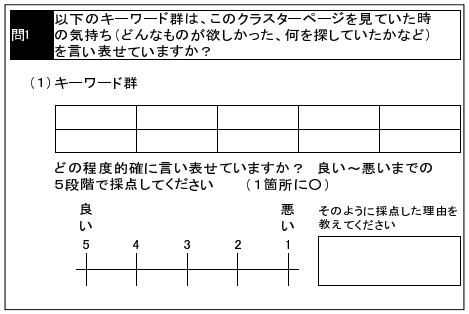

図 6 質問用紙

ウェブ閲覧は閲覧サイトや利用目的に依存して行われ，異なる性質の閲覧履歴を生成する．そのため，評価実験はいずれかの閲覧形態に限定して適用性を検証することが必要である．今回の評価実験における，被験者は F1 層 (20〜30 代女性)12 名に対して行った．各被験者は 1 時間の閲覧セッションを 2 回行い，それぞれの 1 時間のセッションにおいて最低 1 つの商品を指定された EC サイト (楽天，Yahoo ショッピング，Yahoo!オークション，Amazon.JP，ベルメゾン，アットコスメ，ニッセン) のいずれかにおいて，購入することを制約とする．購入の制約は，閲覧が無目的に行われることを回避するためであり，購入サイトを限定しているのはスクレイピングの精度を向上させるためである．被験者の閲覧履歴は 3 章で説明された方式により，処理され，クラスタに与えられたキーワードセットについて閲覧した被験者自身により評価される．

評価対象となるキーワードセットは TFIDF [⋆2]方式により抽出された上位 9 個のキーワード (**方式 1**)，バイグラム連接コスト*IDF 方式により抽出された上位 9 個のキーワード (**方式 2**)，提案方式により抽出された 9 個のキーワード (**方式 3**) の 3 セットである．重みづけ方式として最もよくつかわれる TFIDF と特徴的なキーワードを抽出可能な連接スコア法を比較方式として採用した．

IDF と連接コストの 2 方式はそれぞれの履歴群をコーパスとして学習を行った．また，各ウェブページの長さを正規化するため，提案方式を除く 2 方式は各閲覧履歴における重みの総和が 1 となるよう正規化を行った．

提案方式はコンテキストワード 4 個，ユニークワード 5 個を抽出し，閾値 N を 0.8 と設定した．また，十分な数のコンテキストワードユニークワードが獲得できない場合，閾値 N を 0.05 刻みで下げることでツリーの再作成を繰り返し，閾値 0 の時点でキーワード数が確保できない場合，全てコンテキストワードとして扱い，評価を行った．コンテキストワー

⋆1 回答者が，提示された文 (この場合質問文) へどの程度合意できるかを等間隔の尺度で測る尺度法．
⋆2 IDF 値の算出元となる文書群には処理対象のクラスタ内の全文書を利用した．

Vol.2009-NL-194 No.7
2009/11/16

表 1 各閲覧実験の平均得点の総和

方式 1	76.71
方式 2	75.35
方式 3(提案方式)	74.46

表 2 コンテキストワードとユニークワードが確保できたデータにおける各閲覧実験の平均得点の総和

方式 1	36.83
方式 2	35.83
方式 3(提案方式)	40.33

ドリストの獲得には，x を 20 と設定し，処理対象の直前の 20 履歴のうち，いずれかのクラスタに属していて，かつ，最も新しい履歴が属するクラスタに含まれるキーワードをコンテキストワードとした．

図 6 は質問紙のサンプルである．各方式のキーワードセットは 9 個取得し，ランダムに配置される．キーワードセットの表示順は，質問クラスタごとに，方式 1 →方式 2 →方式 3，方式 1 →方式 3 →方式 2，方式 2 →方式 1 →方式 3，方式 2 →方式 3 →方式 1，方式 3 →方式 1 →方式 2，方式 3 →方式 2 →方式 1，の順に変更した．

また，5 個以下または 26 個以上で構成されるクラスタへの評価セット，一つ以上の評価が記入されていなかった評価セットに関しては外れ値として評価値から除外した．

4.2 実験結果と分析

実験で得られた結果について，表 1，表 2 に示す．表 1 は各閲覧実験ごとのリッカート尺度の平均値を算出し，その総和を比較している．表 2 は提案方式が十分な数のコンテキストワード，ユニークワードが確保できたデータ (コンテキストワード 5 個，ユニークワード 6 個以上が確保できていたデータ) について，各閲覧実験ごとのリッカート尺度の平均値を算出し，その総和を比較している．

総合的な得点については，TF*IDF，連接コスト*IDF，提案方式の順で高い値を示した．しかし，実験全体においてコンテキストワードが少ないため，提案方式が機能していないケースが多く存在していた．そのため，表 2 のように十分な数のコンテキストワード，ユニークワードが確保できたデータのみを対象として再評価を行ったところ，提案方式の得点が他方式より高い値を示した．なお，全実験データ数は 116 データであり，そのうち提案方式が十分な数のコンテキストワード，ユニークワードが確保できたデータ数は 16 データである．

この実験結果より，コンテキストワードとユニークワードを確保できれば，提案方式のフレームワークが有効である可能性を示した．

5. 議　　論

実験を通して得られた知見を以下に示す．

それぞれの方式で抽出されるキーワードの特徴は以下の通りである．TFIDF はファッションやコスメといった抽象度の高い語を抽出することが多く，大きな失点はなかったが，閲覧の詳細を把握するには不足していた．今回の閲覧においては，閲覧サイトを限定することで，ストップワードが有効に機能したため，高い評価値が得られた．LRIDF，提案方式は商品名やブランド名など抽象度の低い語を抽出することができたが，抽出されたキーワードが誤っていた場合，抽象度が低いほど印象付けられる傾向にあり，評価値は TFIDF とほぼ同じ評価値にとどまった．評価データが提案方式の仮説に基づいた閲覧である場合，提案方式は高い評価値を得たが，抽象度を変化させながら行われる閲覧（目的を徐々に絞っていく閲覧など）は全体の一部であり，提案方式の仮説が適用できない閲覧（強い目的を有さない閲覧）に対しては機能していなかった．

また，被験者評価で抽出されるキーワードの半数以上はカテゴリ，ショップ名，ブランド名，検索語が占めており，ウェブにおける購買行動においてはこれらの語に重みを与えることで，有効に機能することが考えられる

表 1，表 2 において，t 検定を用いて平均の検定を行った．検定の結果，表 1，表 2 ともに有意水準 5%では有意差を検出することはできず，有意水準 25%において有意差を検出した．一般にこの水準では有意差があると断定することは困難である．また，評価値の分布は正規分布を描いておらず，評価自体も極端なスコアを付ける被験者と平均値近くを付ける被験者が分かれており，評価方法と検定方法の改善は今後の課題となっている．

6. ま と め

本稿では，ユーザの閲覧意図を考慮したキーワード抽出方式を提案した．また，提案方式とクラスタリング技術と組み合わせたフレームワークを提案し，被験者実験を通してその妥当性を検討した，実験の結果，予期していた有効性は得られなかったが，5〜25 履歴で構成される，コンテキストワードとユニークワードが確保できる，といった条件を満たすことができれば，既存方式より有効に機能する可能性を示した．

参 考 文 献

1) G. Salton, Automatic Text Processing, AddisonWesley,MA.
2) H. Nakagawa, and T. Mori,A simple but powerful automatic term extraction method. In COMPUTERM 2002: Second International Workshop on Computational Terminology pp. 1-7, 2002.
3) Katerina T. Frantzi and Sophia Ananiadou. Extracting nested collocations. In COLING ’96, pp. 41 . 46, 1996.
4) Katerina T. Frantzi and Sophia Ananiadou. The c-value/nc-value method for atr. Journal of NLP, Vol. 6, No. 3, pp. 145 . 179, 1999.
5) S. E. Robertson, H. Zaragoza, and M. Taylor, Simple BM25 extension to multiple weighted fields, in Proceedings of the Conference on Information and Knowledge Management (CIKM), 2004.
6) Mani and E. Bloedorn, Machine learning of generic and user-focused summarization. In Proc. of AAAI-98, pages 821-826, 1998.

7) ユーザの要約要求を反映するためにユーザとのインタラクションを導入した複数文書要約システム, 知能と情報 : 日本知能情報ファジィ学会誌, pp.265-279, 2006.
8) 大澤幸生, ネルス E. ベンソン, 石塚満: KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会誌, Vol. J82-D-I,No. 2, pp. 391.400 (1999).
9) 松尾豊, 石塚満. 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム. 人工知能学会論文誌, Vol. 17, pp. 213.227, 2002.
10) 松尾豊, 福田隼人, 石塚満, ユーザ個人の閲覧履歴からのキーワード抽出によるブラウジング支援, 人工知能学会論文誌 18 巻 4 号, pp.203-211, 2003.
11) Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In WSDM '08: Proceedings of the international conference on Web search and web data mining, pages 251-260, New York, NY, USA, 2008. ACM.
12) Carol C. Kuhlthau. " Inside the Search Process: Information Seeking from the User 's Perspective ". Journal of the American Society for Information Science. vol. 42, no. 5, 1991, p. 361-371.
13) レボウィッツ紀子, 松村敦, 宇陀則彦, 著者とキーワードの関連性に着目した研究領域ブラウジングシステムの試作, Vol. 17, No. 2, pp.75-80, 情報知識学会誌, 2007.
14) 本文抽出モジュール ExtractContent, http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html
15) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.
16) 専門用語自動抽出用 Perl モジュール"TermExtract", http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
17) 長野翔一, 高橋寛幸, 中川哲也: ユーザの要求変化に着目したウェブ閲覧履歴の分類方式, 情報処理学会研究報告. 自然言語処理研究会報告, Vol.2008, No.90, pp.65-70, 2008.
18) 山田 和明, 中小路 久美代, 上田 完次, Web ユーザの行動履歴解析のためのデータマイニング, 電子情報通信学会ヒューマンコミュニケーショングループ WI2 研究会資料, pp.59-64, 広島, Sep, 2005.
19) M.Salton, M.J.McGill, Introduction to Modern Imformation Retrieval, McGraw-Hill, 1983.