A2-1

Web アクセスログのクラスタリングによる 問合せ拡張に関する検討 : i タウンページ上での実験

大浦　勇亮[†]　　喜連川　優[†]

† 東京大学生産技術研究所

E-mail: †{ohura,kitsure}@tkl.iis.u-tokyo.ac.jp

概要　Web サイトを訪れたユーザの行動履歴はアクセスログとして記録しておくことが出来る。ログには、リクエストされた URL アドレス、アクセス日時、リモートホスト名をはじめとして、Cookie やフォームに入力された内容等を残すことが可能である。近年、サイト構築者がサイト構造の見直しを図ったり、サイトに訪れたユーザの行動支援を目的とした、アクセスログの分析・利用に関する研究が盛んになってきている。本研究では、職業別電話番号情報サイトの Web アクセスログを用いて、同一ユーザセッションにおけるユーザの番号情報の問合せに関して、ユーザセッションのクラスタリングを行った。また、得られたクラスタを利用して、ユーザの行動支援を目的とした 1 つのアプリケーションとして、ユーザの問合せ拡張手法を提案した。提案手法では、単に関連性の高い業種を推薦するだけでなく、業種分類上異種性が高いもののログから強い関連性が見出される業種も推薦する。提案手法に基づくシステムを実装し、評価実験により有効性を明らかにした。

キーワード　データマイニング, 知識発見, Web アクセスログ, クラスタリング

Examination of the expansion of users' search queries based on clustering web access logs : Experiments on the i-townpage.

OHURA YUSUKE[†] and KITSUREGAWA MASARU[†]

† Institute of Industrial Science, The University of Tokyo

E-mail: †{ohura,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　Users' access patterns are extracted from the web servers' access logs which consist of URL, access time, IP address, cookies, the contents inputted into forms and so on. Studies on knowledge discovery of web access logs have become important to improve the sites' organization and help users to navigate and get informations they wants. In this paper, we describe results of clustering user sessions and we propose the method to expand the users' search queries based on the clustering results. Our method includes suggestions of related categories although they are non-similar in existing category hierarchy as well as recommendations of similar categories. We also developed a prototype and evaluate it through some experiments.

Key words　Data mining, Knowledge discovery, Web access logs, Clustering

1. はじめに

Webサイトを訪れたユーザの行動履歴はアクセスログとして記録しておくことができる。ログには、リクエストされたURLアドレス、アクセス日時、リモートホスト名をはじめとして、Cookieやフォームに入力された内容等を残すことが可能である。近年、大容量2次記憶装置の低価格化ならびにプロセッサ性能の向上に伴い、膨大なWebサイトに対するユーザのアクセスログ（いわゆるクリックストリーム）を解析することが可能となり、Webログマイニング技術が注目を集めている。データマイニング手法によりアクセスログを分析し、ユーザのアクセス支援や、サイトの再設計のためのツールなど、種々の利用が模索されている。筆者らの知る限り、大規模なログマイニングに関する実験結果の報告は殆どなされていない。

本研究では、職業別電話番号情報サービスを提供する大規模商用サイト[1]のWebアクセスログを用いて、ログマイニングによる問合せ拡張に関する実験結果を報告する。ユーザのアクセスログに対し、K-means法を拡張したクラスタリング手法を用いて分析を行い、その結果を利用した問合せ拡張手法を提案する。提案手法では、分析によりユーザの問合せと類似度が高い業種を推薦するだけではなく、業種分類上異種性の高いものの強い関連性がログから見出される業種の推薦も行う。提案手法に基づくシステムを実装すると同時に、その有効性に関する評価結果についても報告を行う。

以下では、第2章にて関連研究を紹介し、第3章にてアクセスログのクラスタリング手法とその結果を記し、第4章にてクラスタリング結果を利用した問合せ拡張手法を提案、システムの実装、評価結果を述べる。そして第5章にてまとめる。

2. 関連研究

データマイニング手法等を利用したアクセスログによるWebページナビゲーションの分析は盛んに行われているものの[2]〜[6]、分析結果に基づくユーザ支援に関する研究は必ずしも多くなく、小規模実験によるWebページの推薦、リンクの生成程度しか行われていない。Yan等[7]は、ユーザセッションをそのアクセスしたWebページとそのアクセス数で表現してクラスタリングを行い、動的にリンクを生成するシステムを構築した。また、Mobasher等[8]は、ユーザのWebページナビゲーションに関して、Webページ間の相関ルールを抽出し、それを用いてユーザにWebページを推薦するシステムを提案している。

また、アイテムの推薦では協調型フィルタリングの研究が盛んに行われている[9]〜[12]。しかしながら、協調型フィルタリングは、アイテムに対する評価をユーザプロファイルとして保持し、それを利用することが前提となっている。本論文で取り扱うサイトは一般にユーザを特定することは容易ではなく、ユーザプロファイルがない状況下における不特定ユーザに対する問合せ拡張の実現を目的としている。

3. Webアクセスログのクラスタリング

この章では、Webアクセスログのクラスタリング実験に関して、使用したデータについて述べ、その後、ユーザセッションの定義、クラスタリングアルゴリズム、実験結果について記述する。

3.1 実験用データ

今回の実験では、ＮＴＴ番号情報株式会社の協力の下、iタウンページ[1]のWebアクセスログを使用した。iタウンページは、日本全国約1,100万件の店舗情報の検索サービスを提供している大規模商用Webサイトである。サイトには、1ヶ月あたり5000万ページビュー(2001年12月現在)を越えるアクセスがある。ユーザはタウンページ（職業別電話帳）と同じように、店舗名や業種で電話番号情報を調べることができる。図1の左側のページがそのトップページであり、検索を行うには、条件入力フォームからキーワードまたは業種名または店舗・企業名と地域（住所）を指定する必要がある。図1の右側のページが結果のページであり、検索結果には、入力条件に合致する店舗・企業名と住所、電話番号、地図、詳細情報等が出力される。業種の指定では、直接入力することが可能である他、50音別の索引や階層構造になっている業種リストを用いて業種を選択することができる。この業種階層は、図2に示すような階層構造をとっており、ユーザが業種リストから”ホテル”を選択できる画面に到達するには、「レジャー産業」→「旅館・ホテル業」と階層を辿る必要がある。

iタウンページのアクセスログは、通常のWeb

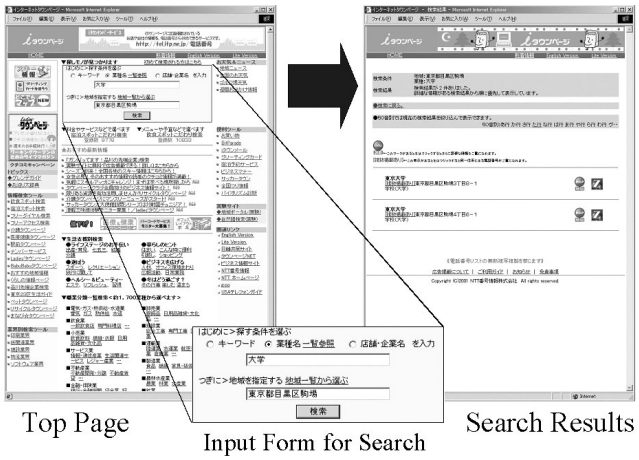

図1　i タウンページのトップページと検索結果ページ

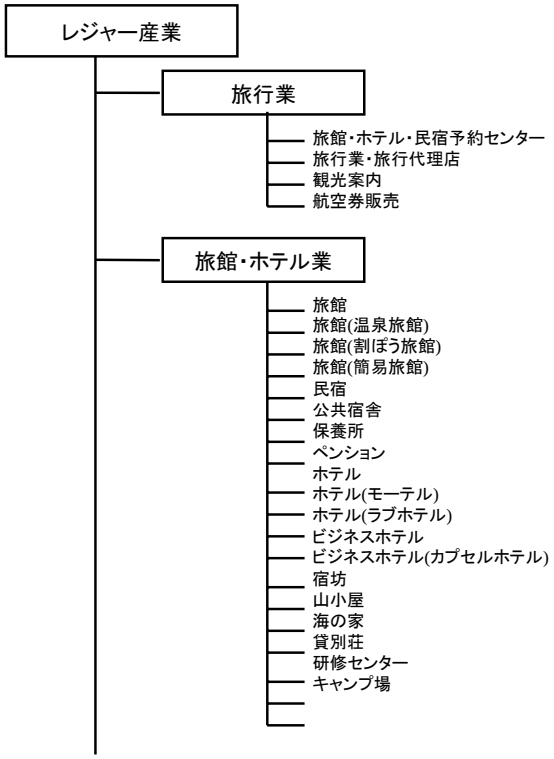

図2　i タウンページにおける業種階層の例

サーバのログに加えてアプリケーションサーバのログから構成されている。これらのログからとれる情報には、アクセス日時、リクエスト URL、リモートホスト名、Cookie ID をはじめとして、入力された検索条件、検索結果件数等がある。今回、分析に使用したデータは 2000 年 2 月 1 日から 6 月 30 日までのアクセスログであり、総ログ数は 450,044,489 リクエストであった (画像、検索以外のリクエストを含む)。このうち、クラスタリング対象としたのは、Cookie ID がとれたユーザの検索リクエスト 24,629,517 件である。

3.2　ユーザセッション

ログには不特定多数の人によるリクエストが全て記述されるが、Cookie ID が取得できている場合は、それによって同一ユーザのリクエストをログから特定することができる。同一ユーザでも訪れる度ごとにその要求は変化するため、通常、アクセスログの分析では、セッションという概念を導入する。セッションは、同一ユーザのリクエストにおいて、各リクエスト間隔が 30 分以内であるリクエストの集合によって表される [13]。本実験では、このユーザセッションを単位として、その問合せに入力された業種の類似するユーザセッションをクラスタリングすることによって、ユーザの問合せ傾向を理解することを目的とする。

3.3　クラスタリングアルゴリズム

まず、ユーザセッションをその入力した業種によってベクトル $\vec{s}$ で表す。ここで、業種総数を N_g とすると、i 番目のセッション S_i のセッションベクトル $\vec{s_i}$ は、式 (1) に示すように、そのセッションにおいて業種 j が入力された場合に 1、入力されていない場合は 0 を値にとる N_g 次元のベクトルとした。

$$s_{ij} = \begin{cases} 1 & \text{業種 j が入力された場合.} \\ 0 & \text{業種 j は入力されていない場合.} \end{cases} \quad (1)$$

クラスタリングアルゴリズムは、一般的に広く使われている K-means アルゴリズムであるが、生成すべきクラスタ数が未知であるため、初期値としてクラスタ数 K を与えるのではなく、類似度閾値 TH_{sim} を与えることによりクラスタ数を動的に決定可能であるように改良を施した。また、類似度の計算には内積 (コサイン) を使用した。アルゴリズムの詳細は以下の通りである。

入力セッション数が N である時、入力ベクトル $\vec{s_1}, \vec{s_2}, \vec{s_3}, .., \vec{s_N}$ に対して、

1. 最初の入力ベクトル $\vec{s_1}$ をクラスタ C_1 の中心ベクトル $\vec{c_1}$ とし、S_1 を C_1 のメンバとする。

2. 以後、入力 $\vec{s_i}$ に対して、既存のクラスタ C_1 ... C_k との類似度を式 (2) によって計算し、どのクラスタとの類似度も閾値 TH_{sim} 未満の場合は、新たなクラスタを生成してそのクラスタ中心とし、類似度が閾値 TH_{sim} 以上の場合は、最も類似度が高いクラスタのメンバとする。この時、メンバが新たに増減したクラス

タはその中心ベクトルを式 (3) にて再計算する。

3. 割当てが収束するまで繰り返す。

$$SIM(i,j) = \frac{\vec{s_i} \cdot \vec{c_j}}{|\vec{s_i}||\vec{c_j}|} \quad (2)$$

$$\vec{c_j} = \frac{\sum_{S_i \in C_j} \vec{s_i}}{M_j} \quad (3)$$

ここで M_j はクラスタ C_j のメンバ数である。

3.4 クラスタリング結果

ユーザセッションのクラスタリングを行うにあたって、2 つ以上の業種が入力されたユーザセッションを対象とした。この時、該当するセッション数は 564,355 であった。得られたクラスタには、メンバ (セッション) 数が 1 であるクラスタも含まれる。得られた結果を解釈するにあたって、小さいクラスタは考慮しないこととし、最小クラスタサイズ MIN_{cl} よりも少ないメンバ数のクラスタはクラスタとしてカウントしないことにする。今回の実験では、この値を対象セッション数の 0.01% (MIN_{cl} = 56.4 人) 未満とし、それに相当する 56 人以下のユーザセッションで構成されるクラスタは切り捨てた。また、いくつかのパラメータで試験した結果、閾値 TH_{sim} には 0.10 を用いることにした。その結果、得られたクラスタ数は 826 個であり、クラスタのサイズは、表 1に示すように、57 セッションで構成される小さいクラスタから 21,029 セッションで構成される巨大なクラスタまで抽出された。この時、構成セッション数の平均値は 678.8、メディアンは 330 であった。

表 1 クラスタサイズ

メンバ (セッション) 数	クラスタ数
[57 , 1000)	667
[1000 , 2000)	106
[2000 , 3000)	34
[3000 , 4000)	8
[5000 ,)	11

クラスタリングを行った結果のうち、“ ホテル ” が含まれるいくつかのクラスタを例として表 2に掲載する（その他のクラスタ例も付録に掲載してある）。結果のクラスタは、そのクラスタを構成するセッション (メンバ) 数、および、クラスタを構成するユーザセッションにおいて入力された業種とその業種を入力したセッション (メンバ) 数で表されている。この時、表示する業種は、新たに設定したクラスタ内閾値 TH_{cat} 以上のセッションが入力したものに限定してある。ここでは、クラスタ内閾値 TH_{cat} をクラスタ構成セッション数の 10% とした。例えば、表 2において、クラスタ 1 は 15318 セッションから構成されるクラスタであり、そのうち “ ホテル ” を入力したセッション数が 15318、“ ビジネスホテル ” が 13654、その他の業種は入力したのがクラスタ構成セッション数 15318 の 10% 未満、つまり 1531 以下であることを表している。

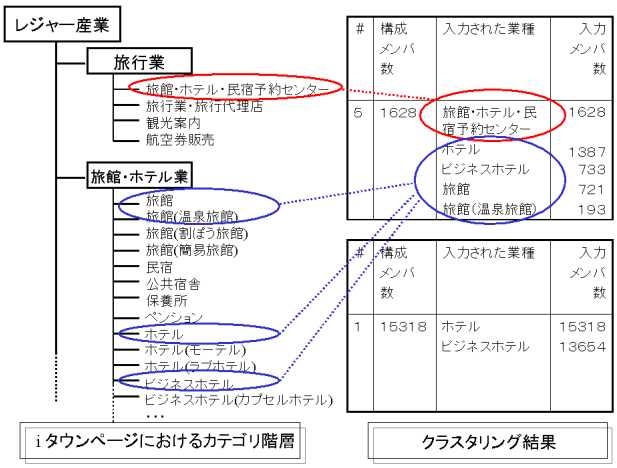

図 3 iタウンページの業種階層とクラスタリング結果

得られた結果のクラスタと i タウンページで用いられている業種階層とを比較すると、業種階層において同一階層にない業種が同一クラスタに存在する結果が数多くみられた。例えば、図 3には左側に i タウンページにおける業種階層の一部が、右側にクラスタリング結果の一部が示してある。この図において、右下のクラスタ 1 は “ ホテル ” と “ ビジネスホテル ” で構成されるクラスタであり、これら 2 つの業種は、カテゴリ階層において両者とも「旅館・ホテル業」というカテゴリの下に位置する同一階層の業種であるが、右上のクラスタ 5 では、その構成業種である “ 旅館・ホテル・民宿予約センター ” は「旅行業」に属する業種であり、その他の構成業種である “ ホテル ”、“ ビジネスホテル ”、“ 旅館 ”、“ 温泉旅館 ” の属する「旅館・ホテル業」とは異なっている。業種 “ ホテル ” が含まれるいくつかのクラスタ例 (表 2) においても、“ 結婚式場 ”、“ レンタカー ”、“ ゴルフ場 ”、“ 貸会議室 ” 等、“ ホテル ” とは業種階層において離れている業種とともに検索に用いられていることがわかる。表 2および付録に掲

表2　クラスタリング結果例

#	構成メンバ数	入力された業種	入力メンバ数	#	構成メンバ数	入力された業種	入力メンバ数
1	15318	ホテル	15318	6	1258	結婚式場	1258
		ビジネスホテル	13654			ホテル ▲	609
2	3293	旅館（温泉旅館）	3293			会館 ▲	303
		温泉浴場 ▲	1153	7	1158	レンタカー	1158
		ホテル	1145			ホテル ▲	120
		温泉供給 ▲	549	8	811	ゴルフ場	811
3	2847	民宿	2847			ホテル ▲	155
		旅館	2448	9	799	スキー場	799
		ホテル	899			ホテル ▲	126
		ビジネスホテル	700	10	732	貸会議室	732
		旅館（温泉旅館）	295			ホテル ▲	215
4	1805	会館	1805			公民館・集会場 ▲	94
		貸会議室 ▲	719	11	346	結婚披露宴演出	346
		ホテル ▲	331			結婚式場	300
		公会堂・会館 ▲ *	211			会館 ▲	104
5	1628	旅館・ホテル・民宿予約センター	1628			祝儀用品 ▲	51
		ホテル △	1387			ホテル ▲	42
		ビジネスホテル △	733				
		旅館 △	721				
		旅館（温泉旅館） △	193				

載したクラスタには、クラスタを構成するトップの業種（上記クラスタ5では“旅館・ホテル・民宿予約センター”）と比較して、その他の業種に関して、その業種が同一階層でない場合、「旅行業」と「旅館・ホテル業」のように親階層まで考慮すれば同一階層である場合は、その業種名の右に“△”を記し、1つ上の親階層まで考慮しても同一階層にない場合にはその業種名の右に“▲”を記した。また、全ての業種が階層リストに分類されているわけではないため、リストに含まれていない業種に関してはその業種名の右に”＊”を記した。

以上より、Webアクセスログのクラスタリングは、表2の結果（および付録の結果）を見てもわかる通り、同じ“ホテル”を入力するにしても、宿泊場所を探している人もあれば、“結婚式場”を探している人、“レンタカー”を探している人、“会議室”を探している人と、ユーザの要求やコンテクストが異なっており、どのような要求の人が存在しているのか、1つのセッションではどういった業種が一緒に入力されているのかを把握することができ、ユーザの行動理解に有効であることが確認された。また、同一階層の業種のみで構成されるクラスタは、826クラスタ中134個(全体の16.2%)、異なる階層の業種を含むクラスタは692個(全体の83.8%)であり、カテゴリ階層では近くに存在していない業種が同一セッションで入力される場合が多いことがわかった。このことは、問合せを行う際に、サイト構築者の意図と異なる意図を持つユーザが存在する可能性があることを示している。

4. 問合せ拡張支援

Webサイト構築者にとって、サイトを構築するにあたって、どの情報(Webページ等)をどこに配置すべきか、どのような階層構造(Webページのリンク階層等)にするのがユーザにとって利用しやすいのかを考えて構築することは非常に重要なことである。しかしながら、これは容易な作業ではない。iタウンページ[1]を利用して、企業や店舗の情報を探すユーザにとって、時として業種を選択するのが困難である。それは、ユーザの意図とサイト構築者の意図の不一致やユーザ要求の多様性に因る。前節にて述べたクラスタリング結果において、異なる階層の業種が同時に入力されることが多いことからもこの問題が生じていることが推察される。

また、検索条件に合致するものが得られない状況に遭遇し、ユーザが当惑する状況もある。iタウンページを利用するユーザにとって、入力した条件にあう結果がなかなか見つけられない状況はしばしば起こり得る。iタウンページのアイテム（企

業番号情報）は、場所と時間の属性を持っているが、場所によっては指定した業種に合致するものがない状況（検索結果 0 件）も数多く存在し、それを知らないユーザは、検索場所の変更を試みた結果、途中で断念する場合も少なくない。 2000 年 2 月から 6 月のログでは、業種と場所による検索リクエスト総数 12,837,141 件に対して、その 30% にあたる 3,888,851 件のリクエストが検索結果 0 件となっている。

そこで、アクセスログのクラスタリング結果を利用して、ユーザがサイトを利用するのを支援するアプリケーションの実現に関して検討する。具体的には、ユーザの問合せに関して入力すべき業種名を拡張支援するシステムの構築を目的とする。例えば、“ホテル”を検索しているユーザに対して、ログ分析から大きく業種は異なるものの“レンタカー”と関連度が高いと判定された場合、レンタカーを関連情報として提示することにより、ユーザにとっての利便性の向上が期待される。また、結果件数が少ない場合、例えば、そもそもその地域には“酒屋”が存在しないのに酒屋を探している場合、問合せ拡張によって、もし“コンビニエンスストア”ならばあるといった状況において、コンビニエンスストアを提示することができ、指定した業種とは異なるものの結果を得られるならば、ユーザの満足度の向上が期待できる。以下では、我々が考案した問合せ拡張システムについて、その問合せ拡張方法と具体的な結果例および評価を述べる。

4.1 問合せ拡張方法

ユーザが問合せ時に入力した業種に対して、問合せ拡張を実現するために、アクセスログのクラスタリング結果と i タウンページにて導入されている業種階層リスト (図 2) を利用した。提案する問合せ拡張方法は 1 次拡張と 2 次拡張の 2 ステップから構成されており、1 次拡張ではクラスタリング結果のみを使用し、2 次拡張ではクラスタリング結果と業種階層リストの両方を使用する。問合せ拡張方法の詳細を以下に示す。

1. 入力された業種が属するクラスタのうち、クラスタ構成メンバ数に対して、その業種を入力したメンバ数の割合が最も高いクラスタを利用して、そのクラスタのメンバが入力した他の業種を提示する。

2. 入力された業種が属するクラスタのうち、1 次拡張で使用していないクラスタにおいてトップにある業種と、サイトに導入されている業種階層とを比較して、入力された業種と同一階層になければその業種を提示する。

1 次拡張では、アクセスログに基づいて、業種を変更する際に使用する可能性が高い業種を提示する。例えば、“ホテル”が入力された場合、“ホテル”の割合が最も高いクラスタ（表 2のクラスタ 1）を用いて、“ビジネスホテル”を提示する。今回の実験では、割合の最も高いクラスタのみを使用したが、拡張範囲を広げたい場合は上位クラスタをいくつか使用すれば良い。

2 次拡張では、さらに、ユーザの意図とは離れている可能性があるものの根拠がある業種を提示する。これは、ユーザによっては非常に有用である場合もあれば、そうではない場合もあり得る。それゆえ、関連度の低い拡張が過度に行われるのを防ぐために使用するクラスタサイズの閾値 MIN_{exp} を新たに導入し高く設定する。表 2を例にとって考えると、“ホテル”が入力された場合、1 次拡張で使用していないクラスタのうち、“ホテル”が含まれるクラスタのトップカテゴリである“旅館（温泉旅館）”、“民宿”、“会館”、“旅館・ホテル・民宿予約センター”、“結婚式場”、“レンタカー”、、、と i タウンページで用いられている業種階層とを比較して、ホテル”と同一階層にない業種である“会館”、“旅館・ホテル・民宿予約センター”、“結婚式場”、“レンタカー”、、、を提示する。

4.2 問合せ拡張結果

前節で述べた構築方針に従って、問合せ拡張支援システムを試作した。このシステムでは、通常の検索と同様に業種名と地域を指定することによって、図 4に示すように、右側のフレームに通常の i タウンページにおける検索結果を表示し、左側のフレームに問合せ拡張された業種を提示する。この図では、検索条件として“スポーツ施設”が指定された場合の検索結果が表示されており、1 次拡張によって“プール”、“競技場”が提示され、2 次拡張によって“スポーツ教室”、“公園”、“スイミング教室”が提示されている。

1 次拡張では、入力された業種入力数のクラスタサイズに対する割合の最も高いクラスタのみを今回は使用した。1 次拡張における提示数を増やしたい場合は、上位のクラスタをいくつか使用する

ことによって実現できる。2 次拡張では有用でない可能性も含まれるため、用いるクラスタサイズの閾値 MIN_{exp} を新たに設定した。今回使用したのは、対象セッション数の 0.1%($MIN_{exp} = 564.4$ 人) であり、564 人以下のクラスタは 2 次拡張に使用しなかった。また、クラスタで使用する業種の閾値 (クラスタ内閾値 TH_{cat}) は前節同様、1 次拡張、2 次拡張ともにクラスタ構成セッション数の 10% とした。

問合せ拡張結果の例を表 3に掲載する。ここで、3の 2 番目の結果は、上記“スポーツ施設”の入力に対するシステムの問合せ拡張結果例と対応している。この時、入力された業種に対して、拡張された業種が他の階層にある場合、1 つ上の親階層まで考慮すれば同一階層にある場合にはその業種名の右に“△”を記し、1 つ上の親階層まで考慮しても同一階層にない場合にはその業種名の右に“▲”を記した。例えば、“スポーツ施設”は、「教育・文化・スポーツ産業」の下の階層である「スポーツ産業」に属し、1 次拡張によって提示される“プール”、“競技場”も同じカテゴリに属しているが、2 次拡張で提示される“スポーツ教室”と“スイミング教室”は「教育・文化・スポーツ産業」の下の階層である「塾・教室・カルチャースクール」のカテゴリに属しているため、それらの業種名の右には“△”を記してあり、“公園”は、「レジャー産業」の下の階層である「公園・遊園地」に属しており、“スポーツ施設”とは 1 つ上の親階層まで考慮しても同一階層にない業種であるため、業種名の右に“▲”を記してある。結果より、2 段階の拡張をすることによって、業種階層としては遠いところに存在する業種に関しても問合せ拡張を行うことが可能であるのを確認できた。

4.3 評　　価

2000 年 7 月 1 日から 7 月 20 日までのログを評価実験用のデータとして、提案手法による拡張の有効性を調査した。まず、実験用のデータをユーザセッションに切り分け、問合せの遷移関係”業種 X →業種 Y”を抽出する。例えば、ある 1 つのセッションにおいて”業種 A →業種 B →業種 C”と入力されていた場合、このセッションからは”業種 A →業種 B”、”業種 B →業種 C”の 2 つの遷移関係が抽出される。この時、遷移前の業種に対する拡張に遷移後の業種が含まれる時、これを拡張成功とする。遷移関係の総数を N とする時、拡張成功

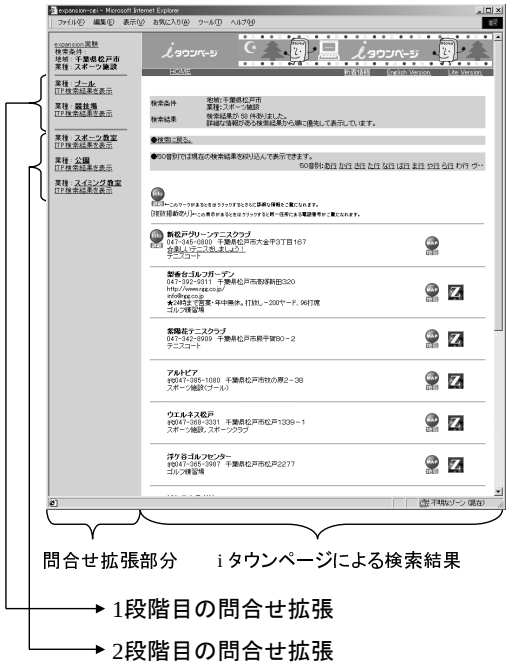

図 4　問合せ拡張支援システムの出力結果

の総数を S とし、拡張成功率を以下のように定義する。

$$\text{拡張成功率} = \frac{S}{N} \tag{4}$$

また、遷移関係 i の入力（遷移関係における左辺）に対する拡張業種数を E_i とすると、拡張業種数の平均は以下のように定義できる。

$$\text{平均拡張業種数} = \frac{\sum_{i=1}^{N} E_i}{N} \tag{5}$$

実験用データでは遷移関係数 N は 318,899 であった。図 5にクラスタ内閾値 TH_{cat} に対する拡張成功率ならびに平均拡張業種数のグラフを掲載する。グラフにおいて、X 軸はクラスタ内閾値を表し、左 Y 軸は拡張成功率、右 Y 軸は平均拡張業種数を表している。

結果より、35% 以上の検索リクエストに対して業種の変更を支援でき、十分に有効であると言える。これより、ユーザが業種の変更を行う手間を軽減できるだけでなく、拡張によって入力すべき業種を発想させる効果も加わり、ユーザの利便性の向上が期待できる。

さらに、問合せ拡張システムでは、検索条件に合致するものが存在しなかった場合に代替案を提示する効果を挙げることもできる。例として、

表 3　問合せ拡張結果例

#	入力業種	拡張（1）	拡張（2）
1	ホテル	ビジネスホテル	旅館・ホテル・民宿予約センター △ 宿泊施設 ▲* 結婚式場 ▲ 貸会議室 ▲ ゴルフ場 ▲ 会館 ▲ スキー場 ▲ レンタカー ▲
2	スポーツ施設	プール 競技場	スポーツ教室 △ 公園 ▲ スイミング教室 △
3	結婚式場	ホテル ▲ 会館 ▲	貸衣装 ▲
4	不動産取引	不動産鑑定 ▲ アパート・マンション ビル管理 △	貸家 △ 駐車場 ▲ 建設業 ▲
5	都道府県機関	都道府県事務所 市区町村機関 警察機関	運輸省 △ 官公庁 △
6	エステティックサロン	美容院 ビューティアドバイザー 化粧品販売 ▲	あん摩マッサージ指圧 ▲
7	葬祭業	ペット霊園・葬祭 霊園	寺院 ▲
8	総合病院	内科	消防署 ▲ 医院 ▲*
9	航空券販売	航空業 ▲	金券ショップ ▲
10	酒屋	酒類卸 ▲ コンビニエンスストア ▲* スーパーストア・マーケット ▲* ディスカウントショップ ▲*	酒造業 ▲
11	コンタクトレンズ	めがね店 ▲* めがね用品製造・卸	眼科 ▲

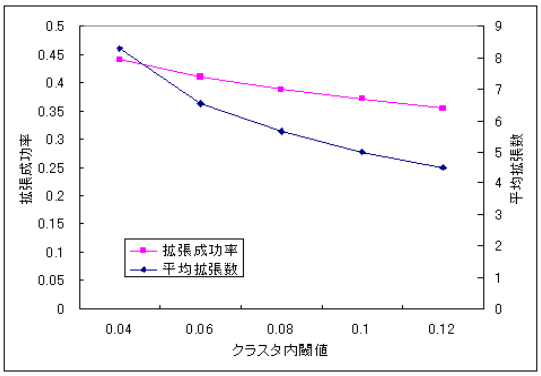

図 5　クラスタ内閾値に対する拡張成功率および平均拡張業種数

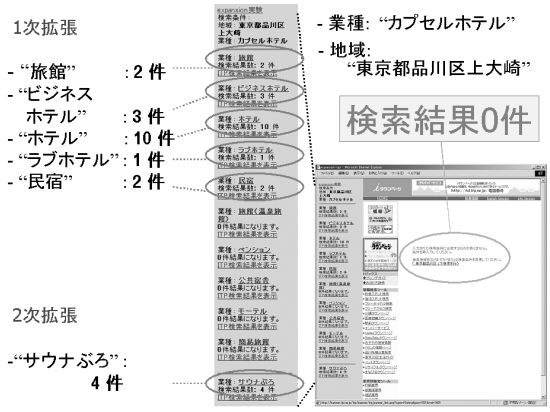

図 6　検索結果 0 件時の問合せ拡張例

図 6に検索結果が 0 件となった場合の問合せ拡張システムの結果を示す。この例では、検索条件に"カプセルホテル"が入力されており、 1 次拡張にて、同じ場所において"旅館"、"ビジネスホテル"、"ホテル"等によってそれぞれ 2 件、 3 件、 10 件、、、の検索結果が得られることを示しており、また、 2 次拡張によって"サウナぶろ"の検索結果が 4 件得られることを示している。

検索結果 0 件時の拡張による結果件数の増分を調べるために、実験データから結果 0 件となる検索条件の業種と場所の組合せを抽出した。図 7に

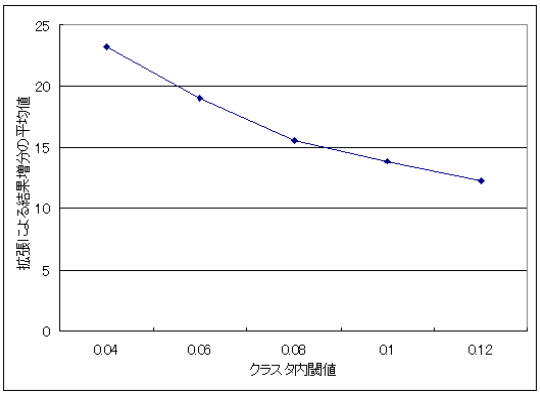

図 7　クラスタ内閾値に対する結果 0 件時の拡張による結果件数の増分の平均値

各クラスタ内閾値に対する 0 件時の業種拡張による結果数の増分の平均値のグラフを掲載する。グラフより、検索結果 0 件において問合せ拡張により少なくとも平均 12 件の結果が得られている。即ち、検索結果に合致しない場合でも拡張によって 12 件以上の代替案を提示することが可能であると期待される。これによりユーザ満足度の向上を図ることができる。

5. おわりに

本稿では、アクセスログを用いて、ユーザセッションにおける問合せに関して、ユーザセッションのクラスタリングを行った。その結果、ユーザセッションにおいてどのような業種が入力されているのかを把握できたと同時に、サイトで導入されている階層構造では距離が離れている業種を入力しているユーザセッションのクラスタの存在が確認された。また、得られたクラスタを利用して、ユーザの問合せを拡張支援する手法を考案した。評価結果より、拡張によって 35% 以上の業種変更に対して有効であり、結果 0 件の検索に対しても関連する他の業種による検索結果を提示でき、提案手法はユーザの利便性の向上に有効であることが示された。

現在のシステムでは 1 入力に対する拡張しか行っていないが、いくつかの入力が与えられた場合、ユーザのコンテクストが推測できる可能性がある。また、時間や場所等によってもユーザの要求が変化する可能性がある。今後はそれらを反映させたシステムの構築を進めるとともに、相関ルールマイニング等他の手法による結果との比較も行っていく予定である。

謝辞　アクセスログを提供していただいているＮＴＴ番号情報株式会社ならびに研究室のメンバーである高橋克己さん、Iko Pramudiono さん、豊田正史さん、Bowo Prasetyo さんをはじめとして、日頃アドバイス、ご協力くださっている方々全てに深く感謝いたします。

文　献

[1] i タウンページ http://itp.ne.jp/
[2] R. Cooley and J. Srivastava and B. Mobasher. Web Mining: Information and Pattern Discovery on the World Wide Web *In Proc.the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.*
[3] C. Shahabe, A. M. Zarkesh, J. Abidi, and V. Shah. Knowledge discovery from user's web-page navigation *In Proc. of the 7th IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE), pp. 20-29, 1997.*
[4] O. R. Zaiane and M. Xin and J. Han. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs *Advances in Digital Libraries 19-29, 1998.*
[5] M. Perkowitz and O. Etzioni. Adaptive Web Sites: Automatically Synthesizing Web Pages *AAAI/IAAI 727-732, 1998.*
[6] O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram. Mining Web Access Logs Using Relational Competitive Fuzzy Clustering *In Proc. of the 8th International Fuzzy Systems Association World Congress - IFSA 99, Taipei, August 1999.*
[7] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking *In Proc. of the 5th International World Wide Web Conference, volume 28, 1007–1014, 1996.*
[8] B. Mobasher and H. Dai and T. Luo and M. Nakagawa. Effective Personalization Based on Association Rule Discovery from Web Usage Data *In Proc. the 3rd ACM Workshop on Web Information and Data Management (WIDM01), 2001.*
[9] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using Collaborative Filtering to Weave an Information Tapestry *Communications of the ACM, 35, no.12,pp.61-70, Dec. 1992.*
[10] P. Resnik, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews *In Proc. of the 1994 CSCW, 1994.*
[11] U. Shardanand and P. Maes. Social Information Filtering: Algorithms for Automating "Word of Mouth" *In Proc. of ACM CHI'95, 1995.*
[12] L. H. Ungar and D. P. Foster. Clustering Methods For Collaborative Filtering *In Proc. of the Workshop on Recommendation Systems. AAAI Press, Menlo Park California, 1998.*
[13] L.D.Catledge and J. E. Pitkow. Characterizing Browsing Behaviors on the World Wide Web *Computer Networks and ISDN Systems, 27(6), 1995.*

付　録

1. Web アクセスログのクラスタリング結果

表 A·1　クラスタリング結果例

#	構成メンバ数	入力された業種	入力メンバ数
1	21029	病院・療養所	21029
		総合病院	19811
		救急病院・救急医療センター ▲*	10055
2	15192	飲食店	15192
		レストラン	2320
3	8886	居酒屋	8886
		飲食店 △	6041
		ろばた焼	913
4	6305	不動産取引	6301
		不動産鑑定 ▲	2413
		アパート・マンション	1583
		ビル管理 △	769
5	4250	ビジネスホテル	4246
		カプセルホテル	1068
6	4146	喫茶店	4146
		コーヒー専門店	676
		レストラン △	456
7	3924	市区町村機関	3894
		町村役場 ▲*	1911
		都道府県機関	634
8	3522	美容院	3522
		理容店	1528
		ヘアーデザイナー	415
9	3328	スーパーストア・マーケット *	3328
		スパゲティ店 ▲	345
10	2887	書店	2887
		古本	752
		書籍販売取次業 ▲	563
		レンタルビデオ ▲	309
11	2774	スポーツ用品店	2774
		スキーショップ	411
		スポーツ用品製造・卸 ▲	393
12	2675	おもちゃ店	2675
		ゲームソフト	375
13	2410	ラーメン	2410
		中華料理店	1048
14	2216	酒屋	2216
		酒類卸 ▲	349
		コンビニエンスストア ▲*	293
		スーパーストア・マーケット ▲*	249
		ディスカウントショップ ▲*	239
15	2215	ホームセンター	2215
		スーパーストア・マーケット ▲*	756
		ディスカウントショップ ▲*	637
		日用品雑貨店	243
16	2131	スポーツクラブ	2131
		スポーツ施設	1530
		プール	502
17	1936	中国料理店	1936
		中華料理店	1869
18	1928	サウナぶろ	1923
		サウナ設備 ▲	870
		銭湯	532
		温泉浴場	500
		カプセルホテル ▲	386
19	1883	ペンション	1883
		民宿	952
		ホテル	769
		旅館	687
		旅館（温泉旅館）	353
		ビジネスホテル	191
20	1610	旅行業	1610
		航空券販売	367
21	1609	携帯電話	1609
		電気通信業	545
		通信用機械器具 ▲	314
22	1417	金券ショップ	1369
		チケット ▲*	907
		ディスカウントショップ ▲*	169
		航空券販売 ▲	156
23	1385	眼科	1385
		コンタクトレンズ ▲	299
		病院・療養所	297
		医院 ▲*	185
24	1366	コンピュータ	1366
		インターネット関連サービス ▲	238
		コンピュータ用品 ▲	186
		ＯＡ機器販売 △	162
25	1355	寺院	1338
		神社	396
		葬祭業 ▲	238
26	1352	エステティックサロン	1346
		美容院	417
		ビューティアドバイザー	229
		化粧品販売 ▲	185
27	1330	図書館	1330
		市区町村機関 ▲	142
28	1080	駐車場	1070
		駐車場（月極）	513
		不動産取引 ▲	246