検索傾向の時間的な類似に基づくクエリ間の関係性判定

小野田　透†　　湯本　高行††　　角谷　和俊†††

† 兵庫県立大学大学院環境人間学研究科　〒 670-0092 兵庫県姫路市新在家本町 1-1-12
†† 兵庫県立大学大学院工学研究科　〒 671-2280 兵庫県姫路市書写 2167
††† 兵庫県立大学環境人間学部　〒 670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †nd07o007@stshse.u-hyogo.ac.jp, ††yumoto@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

あらまし　近年，Web からの情報収集は一般的なものとなり，検索システムやユーザの検索を支援するシステムの重要性は増している．特に，ユーザに対して Web 検索クエリの候補を提示し，検索行動を支援するシステムは Web 検索に不慣れなユーザに対して効果的である．しかし，GoogleSuggest などの従来のクエリ提示システムでは，ユーザが入力したクエリと提示されるクエリの関係性までは考慮されておらず，効率的に検索を行うことができるクエリをユーザが選択することは困難である．本研究では，クエリの検索頻度の時系列データをクエリログより取得し，クエリ間の時間的な検索傾向の類似性を抽出することでクエリ間の関係を判定する手法を提案する．本手法により，クエリ間の関係を考慮したクエリ提示が可能になり，ユーザの検索行動の支援を効率的に行うことができると考えられる．本稿では，提案手法について評価実験を行い，手法の有効性を確認した．

キーワード　クエリログ，時系列データ

Judgement of Relation between Queries Based on Temporal Similarity of Search Tendency

Toru ONODA†, Takayuki YUMOTO††, and Kazutoshi SUMIYA†††

† Graduate School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
†† Graduate School of Engineering, University of Hyogo
2167 Syosya,Himeji,Hyogo,671-2280, Japan
††† School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †nd07o007@stshse.u-hyogo.ac.jp, ††yumoto@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

Abstract　In late years, the Web search has become common practice. Therefore, the Web search system and the search support system has become more important. If a user is unaccustomedness for a Web search, the system that shows the candidate of queries for a user is particularly effective. However, the conventional system do not consider to the relationship of the queries. Therefore a user cannot select the query for search effectively. We proposed the method to judge relations between queries by extracting the similarity of the search tendency of the queries. The system gets time-series data of the search frequency of the query from query-log. In this paper, we conducted experiment to evaluate availability of our proposed method.

Key words　Query-log, Temporal data

1. はじめに

現在，個人の Web からの情報収集は人々にとって日常的なものとなっている．ユーザは必要な情報に関するキーワードをクエリとして Google，Yahoo などの検索エンジンに入力することで，検索結果として提示される Web ページ集合から情報を収集することができる．こうしたキーワードによる検索システムは現在の Web からの情報収集において重要な位置を占めており，多くのユーザがキーワードによる検索システムを用いて情報を収集している．

しかし，現在の検索システムはユーザが自分の必要な情報についてどのようなクエリで検索すれば情報が得られるかを理解しているという前提で設計されている．よって，自分の知りたいことを明確にキーワードとして入力できないユーザは適切に検索を行うことができず，まず自分に必要な情報を調べるためのクエリを作成するために，情報を収集しなければならない，といった問題が発生している．そのような問題を解決するため，Google Suggest [1] のようにユーザの入力に対して，その入力キーワードを含んだクエリを提示するシステムが提供されている．しかし，GoogleSuggest のような従来のシステムで提示されるクエリは，ユーザの入力したクエリとどのような関係にあるクエリなのかという情報は提示されず，ユーザは提示されたクエリによって検索を行い検索結果の確認を行わなければならないなど，あまり効率的な支援であるとは言えない．こうした提示を効率的に行うためには，ユーザの入力したクエリとシステムが提示するクエリの関係性を考慮する必要があり，クエリ間の関係性を求めるための手法が必要であると考えられる．例えば，検索対象を明確にキーワードとして入力できないユーザに対しては，ユーザの入力したクエリについて，検索対象をより具体的にしたクエリを提示するべきである．

本研究ではクエリが「これまでどのような傾向で検索されてきたか」というクエリの時間的な検索傾向を用いてクエリ間の関係性を判定し，判定された関係に基づきクエリを分類して提示することでユーザに対する検索支援を行う手法を提案する．クエリの時間的な検索傾向を用いることにより，ユーザの入力したクエリと提示するクエリ間の関係性を判定することが可能になる．また，本研究では，クエリの検索頻度の時系列データ，過去に検索エンジンに入力されたことのあるクエリのデータを取得するためにクエリログを利用することを想定している．クエリログとは，ユーザが検索に用いたクエリの履歴であり，検索エンジンは日々この履歴を蓄積している．クエリログには多様なデータが含まれるが，本研究で利用するデータは以下の 2 つである．

- 検索エンジンに過去に入力されたクエリ
- クエリの検索頻度の時系列データ

以降，2 節では本研究の関連研究について，3 節ではクエリ間の時間的な類似性の定義と判定方法について述べる．そして，4 節ではクエリの時間的な類似性を用いたクエリの提示システムの提案を行う．5 節で提案手法の評価実験について述べ，6 章でまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 検索支援のためのクエリ提示

ユーザにクエリを提示するための手法として，ユーザの入力した語に対して関連のある語を提示するものが考えられる．関連語を取得するシステムとして Google Sets [2]，Google Suggest というサービスがある．これらのサービスではいくつかの関連語と考えられる語を与えると，それらが含まれるような関連語の一群を抽出し結果として返す．Google Sets のアルゴリズムは公開されていないが，Google が収集した Web ページに含まれる語に対して大規模なクラスタリングを行っているようである．

また，関連語を検索するための電子化された辞書も開発されている．例として，WordNet [3] や言語工学研究所デジタル類語辞典（シソーラス）[4] などがあげられる．これらを用いることで，様々な関連語が取得できる．これらの研究はあくまで語と語の関係を抽出して関連するものを提示するものであり，クエリとしての関係性を用いている本研究とは異なっている．

Rosie らはユーザが過去に入力したクエリのデータを用いて，入力されたクエリと，その後に修正して入力されたクエリの類似度から意味的に同一のクエリ，意味的に近いクエリ，言い換えが可能なクエリの 3 つの関係を定義，判定し，ユーザに提示する手法を提案している [5] [6]．提案の目的において本研究と非常に類似しているが，本研究ではクエリの関係を時間的な検索傾向の類似性を用いて判定しているのに対し，彼らは入力されたクエリ間のフレーズ的な類似性を用いている点でアプローチが異なっている．大島らは，共通の上位語を持つ語を同位語と定義し,「や」という並列助詞で連結されるキーワード同士は同位語であるとして同位語の発見手法を提案している [7]．例えば,「トヨタやホンダ」と「ホンダやトヨタ」が Web 上の文書に多く存在するのならば“ トヨタ ”と“ ホンダ ”が同位語であるという考えである．また，山口らは，クエリログにおいて，同一のキーワードと一緒に入力されるキーワード同士は同位語であるとし，同位語を発見する手法を提案している [8]．これは,｛トヨタ，自動車｝や｛ホンダ，自動車｝というように共通ののキーワードによっクエリが構成されるようなキーワード同士は同位語である可能性があるという考えである．この場合，2 つのクエリ｛トヨタ，自動車｝と｛ホンダ，自動車｝は「自動車」という共通のキーワードを含むことから，キーワード「トヨタ」と「ホンダ」は同位語である可能性がある．さらに，大島らは，Web 検索での同位語発見とクエリログでの同位語発見を併用した手法も提案している [9]．これらの手法ではクエリにおけるキーワードの共起に基づいて関係が抽出されており．キーワード間の時間的な関係は考慮されておらず，本手法とはアプローチが異なる．

2.2 キーワード間の時間関係の利用

クエリログを用いた先行研究として，Chien らはクエリの検索傾向の類似度を相関係数によって計算し，類似するクエリを発見する手法を提案している [10]．また，Wang らは検索に用いられたクエリのアスペクトを用いて，検索結果の分類とラベル付けを行う手法を提案している [11]．Zao，Baeza-Yates らは，あるクエリを入力したユーザがどのような Web ページを閲覧したかによって，クエリの類似性を計算し，Web ページの改善などに用いる手法を提案している [12] [13]．彼らは検索キーワードの時間的な関係は考慮しておらず，本研究はクエリの時間的な検索傾向によってクエリの関係性判定を行う点で異なっている．また，我々は，クエリを入力したユーザが閲覧したページなどの情報は利用しておらず，クエリの検索頻度の時系列データのみを用いて分類を行う．

甲谷らは，クエリログの中から，検索キーワードとその検索

キーワードによって閲覧されたページの URL 情報を用いて，Web ページがどのような検索キーワードによって検索されて，閲覧されているかを解析することで，Web ページが閲覧されるに至る典型的（一般的）なキーワードの発見手法を提案している [14] [15]．この研究では，Web ページに対して，典型的なキーワードを発見することを目的としており，時間的関係性を用いてクエリ間の関係を判定するという目的とは異なる．

3. 時間的類似関係

3.1 関係の定義

本稿では，クエリの時間的な検索傾向の類似性によってクエリ間の関係を判定し，その判定された関係を時間的類似関係と呼ぶ．時間的類似関係とは，クエリの検索頻度の時系列変化に有意な類似点が存在するクエリ間の関係である．関係の判定を行う 2 つのクエリ間において，定義する時間的類似関係が存在するのであれば，2 つのクエリ間には有意な関係が存在すると考えられる．本手法では，まずクエリの時間的な検索傾向の類似のパターンによって 2 種類の関係を定義する．

全体類似　2 つのクエリ p, q の時間的な検索傾向が互いに全体的に類似する関係を，全体類似と定義する．全体類似となるクエリの例として，{秋葉原，殺人} と {秋葉原，加藤} などがある（図 1）．これらのクエリは共に 2008 年 6 月に発生した秋葉原通り魔事件について検索したクエリであると考えられる．全体類似はこのようなクエリの検索が行われた時期，検索頻度の増減のタイミングなどがほぼ一致するクエリ間の関係である．このような関係にあるクエリは，クエリを構成するキーワードが異なっていても，同一の対象を検索すると考えられる．

部分類似　2 つのクエリ p, q の時間的な検索傾向の一部分が類似する関係を，部分類似と定義する．部分類似となるクエリの例として，{サブプライム} と {リーマンブラザース} などがある（図 2）．{サブプライム} は 2007 年の 3 月頃から多く検索され始め，2007 年 7 月頃から現在までは一定数以上検索され続けている．{リーマンブラザーズ} は 2008 年 9 月頃から多く検索され始めており，2008 年 9 月頃の 2 つのクエリの時間的な検索傾向は類似している．このように，部分類似は全体的な検索傾向ではなく，ある一時期の検索傾向に着目したときに検索頻度の増減のタイミングなどがほぼ一致するクエリ間の関係である．また，検索傾向の一部分が一致するというだけではクエリ間の関係を明確にすることは難しいと考えられることから，本研究ではそれぞれクエリが検索されている時期に関して，包含関係が存在しているクエリ間の関係を部分類似としている．ここでの包含関係とは，一方のクエリについて検索が発生している区間が，もう一方のクエリについて検索が発生している区間に完全に含まれている関係を指す．例えば，図 2 の {サブプライム}，{リーマンブラザーズ} は {サブプライム} の検索が発生している区間が {リーマンブラザーズ} 検索が発生している区間を完全に包含する関係になる．このような関係にあるクエリでは，クエリの検索するトピックに概念的な上位下位関係が成立すると仮定する．つまり，クエリ q がクエリ p を包含する $(q \supset p)$ のとき，クエリ p によって検索されるトピックは，クエリ q によって検索されるトピックの下位トピックであると考える．部分類似であるが包含関係を持たないクエリは，互いにある一時期検索傾向が類似しているが，その他の時期ではそれぞれ異なる傾向で検索されている．このような関係にあるクエリでは，一時期同一の対象の検索に用いられたと考えられるものの，クエリの検索対象間の関係は明確には判定できないため，部分類似とは判定しない．

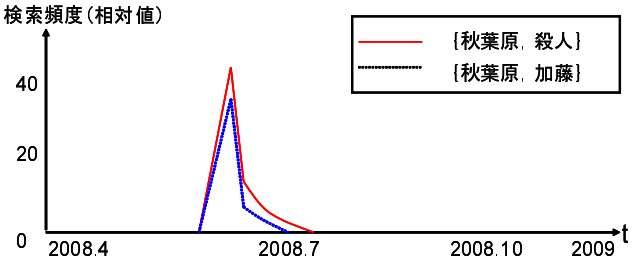

図 1　全体類似の例

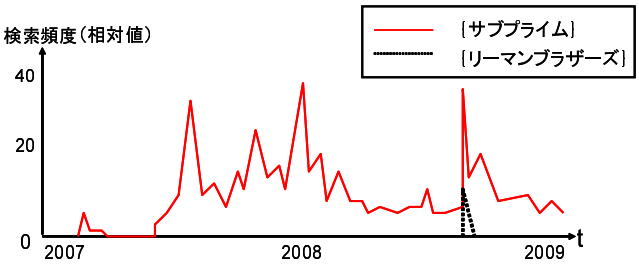

図 2　部分類似の例

3.2 関係の判定

時間的類似関係の判定方法について述べる．ここでは関係の判定を行うクエリをそれぞれクエリ p，クエリ q とする．クエリ間の関係の判定は，以下の手順で行う．

（1）　検索傾向の変化パターンの抽出
（2）　類似区間の抽出
（3）　類似区間を用いた関係の判定

まず，p, q の時間的な検索傾向を検索頻度の増減の変化パターンに変換する．次に，抽出した p, q の検索頻度の変化パターンを比較し，パターンが類似している区間を検索傾向が類似している区間として抽出する．そして，抽出した検索傾向の類似区間によって，時間的類似関係の判定を行う．

- 検索傾向の変化パターン抽出

検索傾向の変化パターン抽出について述べる．本手法では，検索傾向が類似している区間を発見するため，関係の判定を行うクエリ p, q の時間的な検索傾向を検索頻度の増減の時間的な変化パターンに変換する．最初に，全時区間 T を i 個の一定の長さの区間に分割する．本稿では，この区間の長さは 7 日間と設定している．分割した時区間を時間的に過去に存在する区間から順に $t_1, t_2, t_3 ... t_i$ とする．

次に，分割した時区間におけるクエリの検索頻度の比較によって検索傾向の変化パターンを抽出する．各時区間 t_i における検索頻度は，時区間 t_i 中に発生した検索頻度の総和である．クエリ p の時区間 t_i における検索頻度を $qf(p, t_i)$ と表す．変

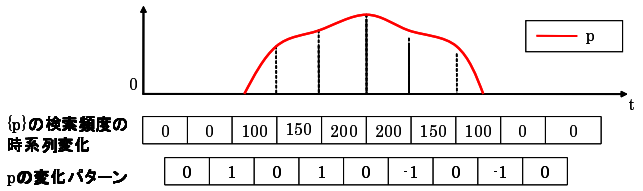

図 3　検索傾向の変化パターン

化パターンの抽出は，時間的に隣合う時区間，即ち時区間 t_i と時区間 t_{i+1} の検索頻度を比較することで抽出される．変化パターンの抽出式を以下に示す．

$$\delta(p,t_i)=\begin{cases}1 & (\text{if } qf(p,t_{i+1})-qf(p,t_i))>\theta_{\text{change}})\\ 0 & (\text{if } |qf(p,t_i)-qf(p,t_{i+1})|\leqq\theta_{\text{change}})\\ -1 & (\text{if } qf(p,t_i)-qf(p,t_{i+1})>\theta_{\text{change}})\end{cases} \quad (1)$$

$\delta(p,t_i)$ は時区間 t_i と時区間 t_{i+1} における変化パターンを返す関数である．$qf(p,t_{i+1})$ - $qf(p,t_i)$ の演算結果が閾値 θ 以上，即ち検索頻度が増加傾向にあるならば 1 を返す．$qf(p,t_i)$ と $qf(p,t_{i+1})$ の差が閾値 θ 以内，即ち検索頻度がほとんど変化しない傾向にあるならば 0 を返す．$qf(p,t_i)$ - $qf(p,t_{i+1})$ の演算結果が閾値 θ 以上，即ち検索頻度が減少傾向にあるならば-1 を返す．これらの値 $\int(p,t_i)$ を t_i におけるパターン値と呼ぶ．ただし，$qf(p,t_i)$，$qf(p,t_{i+1})$，$qf(q,t_i)$，$qf(q,t_{i+1})$ が全て 0 である場合，その区間は意味の無い区間と考える．このような時区間はクエリ p,q が全く検索されなかった時区間であり，p,q の類似性に関係しないと考えるためである．このような時区間をゼロ区間と呼び，以下の式で定義する．

$$Zero(p,q)=\{(t_i,t_{i+1})|qf(p,t_i)=qf(q,t_i)=0,\ qf(p,t_{i+1})=qf(q,t_{i+1})=0\} \quad (2)$$

また，変化パターンの抽出例を図 3 に示す．

- 類似区間の抽出

抽出した検索傾向の変化パターンを用いて，クエリ p,q における類似区間の抽出を行う．類似区間の抽出は，p,q で同一の時間軸に存在する区間のパターン値を比較することで行う．類似区間の抽出は，以下の手順で行う．

（1） 類似区間の始点となる時区間 t_s を定める．t_1，t_2 と順に区間を走査し式 3 を満たす最初の区間を始点とする．

$$\delta(p,t)=\delta(q,t)\neq 0 \quad (3)$$

（2） 終点となる時区間 t_{s+n} を求める．始点 t_s から順に t_{s+1},t_{s+2} と区間を走査し式 4 を満たす最初の区間を終点とする．

$$\delta(p,t)\neq\delta(q,t) \quad (4)$$

（3） t_s から t_{s+n} までの区間を時間的に連続して p，q の検索傾向が類似する区間と見なし，類似区間として抽出する．

上記の手順を繰り返すことによって，p,q 間の検索傾向の変化パターンが類似している区間を抽出する．このような手順によって抽出された区間を部分類似区間と呼ぶ．抽出されたクエリ p,q における部分類似区間を時間的に過去に存在するものから順に，s_1, s_2...s_k とし，部分類似区間の集合を S とする．

- 類似区間を用いた関係の判定

抽出された類似区間を用いて，時間関係の判定を行う．全体類似の判定は p,q パターン値の比較を行った全区間における類似区間の占める割合によって判定される．本稿では，この割合をクエリ間の時間的類似度と呼ぶ．p,q 間の時間的類似度 $Sim(p,q)$ は以下の式によって定義する．

$$Sim(p,q)=\frac{Num(p,q)}{|Periods(p)\cup Periods(q)|} \quad (5)$$

ただし，

$$Periods(p)=\{t_i|qf(p,t_i)>0\} \quad (6)$$

$Num(p,q)$ は p,q 間の部分類似区間の数を返す関数であり $Periods(p)$ は p のみが検索されている区間を表す．$Sim(p,q)$ の値が閾値 θ_{all} 以上であれば，p,q は全体類似であると判定される．

次に部分類似の判定について述べる．部分類似ではクエリは全体的に類似している必要は無いが，クエリの検索区間に対してあまりに類似区間が短いとそれに意味があるかどうかを判定できない．よって，p,q の検索発生区間に対する部分類似区間の最大長の割合 $part(p,q)$ が閾値 θ_{part} 以上であるかの判定を行う．

$$part(p,q)>\theta_{part} \quad (7)$$

ただし，

$$part(p,q)=\frac{Length(p,q)}{|Periods(p)\cap Periods(q)|} \quad (8)$$

$Length(p,q)$ は p,q が共に検索されている区間における s_k の最大長を返す関数である．式 7 を満たす区間が 1 つ以上存在する場合，有意な類似区間が存在すると見なす．

次にクエリ p,q の検索発生区間において包含関係が存在するかどうかの判定を行う．判定は，部分類似区間を用いて行う．包含関係の存在するクエリでは，クエリ p,q における部分類似区間の集合 S が，いずれかのクエリの検索発生区間に完全に含まれていることになる．つまり，クエリ p の時間的な検索傾向にクエリ q が包含される場合 $(p\supset q)$ と，クエリ q の時間的な検索傾向にクエリ p が包含される場合 $(q\supset p)$ がある．よって，以下のような条件で判定を行う．

$$Periods(q)\setminus Periods(p)=S \quad (9)$$

$Periods(q)\setminus Periods(p)$ は $Periods(p)$ が $Periods(q)$ の補集合であることを表す．式 9 は部分類似区間の集合 S が $Periods(q)$ には含まれるが $Periods(p)$ には含まれない，つまり，p，q の部分類似区間が，どちらか一方のクエリの検索区間に完全に含まれる場合に満たされる条件である．式 9 を満たすとき，クエリ p,q 間の関係を部分類似と判定する．

4.　時間的類似関係に基づくクエリ提示システム

提案システムの概要図を図 4 に示す．

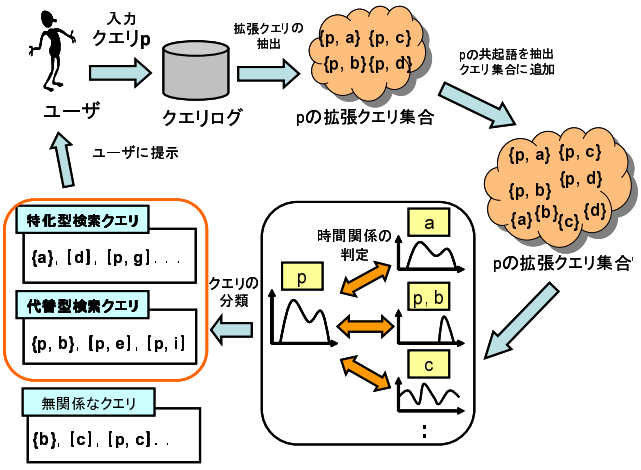

図 4　提案システム概要

提示クエリ間の時間的類似関係を考慮したクエリの提示システムについて述べる．ユーザの入力したクエリとシステムの提示するクエリの関係を考慮しつつ，システムがクエリを提示することで，ユーザの検索行動の支援を行うことが可能になる．本システムでは，クエリを適切に入力することができない場合にもいくつかの種類があると考え，それらに対応するために，ユーザの入力したクエリに対して 2 種類の異なる関係を持つクエリを提示する．

自分の検索目的に対して適切なクエリを上手く入力することができない場合の 1 つは，必要な情報についてのキーワードを明確にできず，曖昧な意味を持つクエリしか入力することができない場合である．例えば，自分はサブプライム問題について，その影響を受けた企業などについて調べたいにも関わらずクエリを明確にできず，{サブプライム} という曖昧なクエリを入力してしまうような場合である．このようなユーザに対しては，入力されたクエリよりも，より検索の対象について話題が明確になったクエリを提示すべきであると考えられる．例えば，上記のユーザに対しては {サブプライム，リーマンブラザーズ}，あるいは {リーマンブラザーズ} というような，入力クエリよりも具体的な話題を特定して検索を行うことが出来るクエリを提示するのが望ましい．本研究ではこのような入力クエリに関する話題をより具体的にして検索することが可能なクエリを特化型検索クエリと定義する．

もう 1 つの場合として，自分の必要な情報について適切に検索を行うことが出来る程度キーワードをいくつかは入力することができるが，それ以外のクエリを思いつかないという場合が考えられる．1 つのクエリから検索される結果だけでは取得できる情報が偏る可能性があり，より多くの情報を得るためにはユーザは検索対象について様々なクエリで多角的な検索を行う必要がある．よって，このようなユーザに対しては，入力したクエリの検索対象について，異なるキーワードをによって検索を行うことが可能なクエリを提示すべきであると考えられる．このような入力クエリに関する話題を異なる観点から検索することが可能なクエリを代替型検索クエリと定義する．本システムでは，2 種類のクエリをユーザに対してその関係を明確にして提示することで，ユーザが状況にあわせてクエリを選択できるようにし，ユーザの検索行動を支援する．

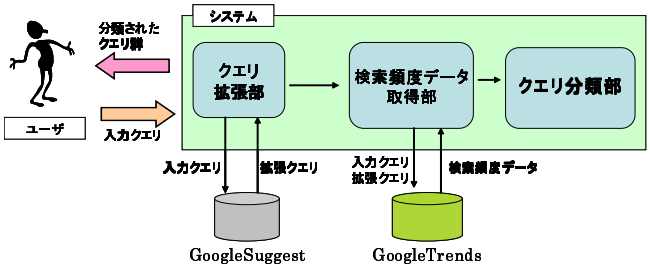

図 5　プロトタイプシステムの構成

提案システムでは，まずユーザの入力したクエリをクエリログのデータを用いて拡張を行い，提示するクエリの候補を取得する．拡張は 2 段階行われ，まずユーザの入力したクエリを構成するキーワードを全て含むクエリをクエリログから抽出する．ユーザの入力したクエリを p とする．p は 1 語以上のキーワードで構成されるクエリである．システムは $\{p, a\}$，$\{p, b\}$ といった，ユーザの入力したクエリにさらにキーワードを付与し，拡張を行ったクエリを抽出する．そして，抽出されたクエリから，ユーザの入力したクエリと共に入力されたキーワードをさらに抽出し，そのキーワードを提示するクエリの候補に加える．そのようにして取得したクエリの集合から検索頻度の高いものから順に n 件を抽出する．

次に，拡張によって得られた拡張クエリ q とユーザの入力したクエリ p との時間的類似関係を判定し，クエリの分類を行う．このとき，p と拡張クエリとの時間的類似関係が全体類似であれば p と拡張クエリは入力クエリ p と同一の対象を検索するクエリ，代替型検索クエリであると判定する．時間的類似関係が部分類似であれば，さらに包含関係であるか部分一致関係であるかの判定を行う．クエリ間の関係が部分類似かつ包含関係 $(p \supset q)$ であるとき，拡張クエリをユーザの入力クエリよりも絞り込んだ対象の検索を行うクエリ，特化型検索クエリであると判定する．時間的類似関係がいずれとも判定されないクエリは無関係なクエリとして判定する．最後に，各関係ごとに分類したクエリ集合をユーザに対して提示し，ユーザの検索行動の支援を行う．

5. 評　　価

5.1 プロトタイプシステム

本システムの構成を図 5 に示す．システムは以下のユニットによって構成される．

- クエリ拡張部
- 検索頻度データ取得部
- クエリ分類部

以下に，各ユニットについて述べる．

- **クエリ拡張部**

入力クエリ拡張部ではユーザの入力したクエリを拡張し，拡張クエリを生成する．クエリログデータの取得が困難であることから，本システムではクエリログの代替として GoogleSuggest

を用いている．GoogleSuggest ではユーザがクエリを入力した場合に，入力したクエリを構成するキーワードを全て含み，過去に検索エンジン（Google）に入力されたことのあるクエリを提示する．本システムでは，GoogleSuggest によって提示されるクエリを拡張クエリとして取得する．また，クエリの拡張として，GoogleSuggest から取得したクエリからユーザの入力したクエリに対して追加されたキーワードだけを抽出し，提示候補クエリとする．

- **検索頻度データ取得部**

検索頻度データ取得部では入力クエリ，拡張クエリの検索頻度の時系列データを取得する．本システムでは，クエリログの代替として GoogleTrends を用いて検索頻度データを取得する．GoogleTrend では，入力したクエリについて 2004 年 1 月以降の検索頻度データを取得することが出来る．GoogleTrends から取得される検索頻度データは，検索頻度の実数ではなく，2004 年以降の検索頻度の増減を相対的な数値で表したデータが csv 形式で取得出来る．提案手法においてはクエリの検索頻度の増減のパターンを抽出することが可能ならば検索傾向の類似性を判定することが可能なため，相対的な数値であっても問題は無い．本システムでは全ての入力クエリ，提示候補クエリをこれらのシステムに入力し，検索頻度の時系列データを取得する．

- **クエリ分類部**

クエリ分類部では，入力クエリと全ての提示候補クエリの検索頻度を用いて入力クエリと提示候補クエリ間の時間的類似性を判定し，判定に基づいてクエリの分類を行う．時間的類似性の判定を行うことで，提示するクエリの決定を行う．

5.2 関係の判定精度の検証

提案手法により，入力クエリと提示クエリとの関係が適切に判定されているかを検証するため実験を行った．検証に用いるデータとして，入力クエリを 10 件準備した．このとき，入力クエリの選定は任意で行った．そして，クエリの拡張手法を用いて，各入力クエリについて 20 件の提示クエリ候補を取得した．ただし，検索エンジンのクエリログの取得が困難であったため，ここでは GoogleSuggest で代用した．GoogleSuggest に入力クエリを 1 件ずつ入力し，提示された 10 件のクエリを提示候補クエリとして取得，さらに拡張手法を用いて計 20 件の提示候補クエリを取得した．これらのクエリの検索頻度の時系列データは GoogleTrends から取得できるデータによって代用した．GoogleTrends では，クエリの検索頻度の時系列データを csv 形式で取得することが出来る．検索頻度データを取得する範囲は 2004 年 1 月 4 日から 2009 年 1 月 11 日までとした．検証に用いた入力クエリを表 1 に示す．

取得したデータを用いて，被験者 5 名に入力クエリとそれに対応する各提示候補クエリの間の関係が以下のいずれに該当するか判定を行った．

（1） 入力クエリと提示候補クエリは，常に同じ対象の検索に用いられる

（2） 入力クエリと提示候補クエリは，同じ対象の検索に用いられる時期がある

（3） 入力クエリと提示候補クエリは全く異なる対象の検索に用いられる

表 1 判定精度の検証に用いた入力クエリ

No	入力クエリ
1	旅行
2	野球
3	サッカー
4	小室哲哉
5	イージス艦
6	耐震偽装
7	サブプライム
8	甲子園
9	オリンピック
10	ワールドカップ

表 2 各関係の判定精度

	正解数/判定数	判定精度
全体類似	3/17	17.6%
部分類似	71/93	76.3%
無関係	25/45	55.6%
合計	99/155	63.9%

(1) は全体類似に，(2) は部分類似に対応しており，(3) は時間的類似関係を持たない，無関係なクエリに対応している．この判定を全ての入力クエリ，提示候補クエリ間に対して行ってもらい，最も回答数が多い回答を正解として，システムの判定結果と照合を行った．ただし，GoogleTrends ではクエリの検索頻度が一定値以上でない場合，検索頻度データを取得出来ない．このようなクエリは判定不可として検証データから除いた．各関係別の判定結果を表 2 に示す．全正解データとシステムの判定の一致率は約 63.8%であった．しかし，全体類似の判定精度が 17.6% と低い値となった．誤判定の多くは，被験者が部分類似と判定したクエリをシステムが全体類似と判定したものであった．つまり，提案システムにおいて全体類似と判定されるクエリでも，クエリを入力するユーザの側からは部分類似，つまり入力クエリとは一時的に関連するクエリであると判定されたケースが多くなった．このことから，本研究で全体類似として定義しているクエリの中にも，部分類似の性質に近いクエリなどが含まれていると考えられ，関係をより詳細に分類する必要があると考えられる．また，提案手法では関係を判定するクエリ間の検索頻度の変化量の類似度は考慮しておらず，一方のクエリの検索頻度が急激に上昇，もう一方のクエリの検索頻度が微増という場合でも，時間的な検索傾向としては類似していると判定している．そのため，本稿における判定手法と検索頻度の変化量を考慮した手法とを比較し，どちらの精度が優れているかを検証する必要があると考えられる．

5.3 システムの有効性の検証

提案手法の有効性について検証するため実験を行った．実験は代替型検索クエリ，特化型検索クエリのそれぞれについて行った．

5.3.1 代替型検索クエリの有効性の検証

代替型検索クエリの有効性を検証するため，実験を行う．代替型検索クエリの有効性は，そのクエリを用いて検索を行うこ

表 3　代替型検索クエリの実験データ

No	入力クエリ	代替型検索クエリ	検索目的
1-1	{ディズニーランド}	{ディズニーリゾート}	ディズニーランドについて様々な観点から情報を得たい
1-2		{ディズニーシー}	
1-3		{東京ディズニーランド}	
2-1	{サッカー}	{サッカー，日本代表}	サッカーについて様々な観点から情報を得たい
2-2		{ワールドカップ}	
2-3		{日本代表}	
3-1	{旅行}	{ツアー}	旅行プランを立てるための様々な情報を得たい
3-2		{旅行会社}	
3-3		{京都}	
4-1	{ジブリ}	{宮崎駿}	ジブリについて様々な観点から情報を得たい
4-2		{スタジオジブリ}	
5-1	{オリンピック}	{五輪}	オリンピックについて様々な観点から情報を得たい

とにより，ユーザの検索の目的を外れずに新しい情報をどれだけ得ることが出来るかによる．よって，代替型検索クエリの検索結果によって，入力クエリの検索結果からは得ることの出来ない新しい情報をどれだけ得ることが出来るかを評価する．また実験では，入力クエリの検索結果の上位 10 件までを閲覧したユーザが，さらに新しい情報を求めて検索行動を行う際に，クエリの変更を行わずに入力クエリの上位 11〜20 位までを閲覧した場合と，代替型検索クエリの上位 10 位までを閲覧した場合を比較し，どちらが有効であるかを検証した．

実験では以下のデータを 1 セットとして用いる．

- 入力クエリ 1 件
- 入力クエリの代替型検索クエリ 1 件
- 入力クエリの検索結果に含まれる URL 上位 1〜10 位の Web ページデータ
- 入力クエリの検索結果に含まれる URL 上位 11〜20 位の Web ページデータ
- 各代替型検索クエリの検索結果に含まれる URL 上位 1〜10 件の Web ページデータ

実験データとして，上記のデータを 12 セット準備した．各入力クエリの選定は任意で行い，代替型検索クエリは入力クエリを GoogleSuggest に入力することで得られたクエリ群の中から抽出した．実験データとして準備したクエリの検索傾向の時系列データは，GoogleTrends によって取得した．各クエリの検索結果に含まれる URL は，Google を用いて検索を行い取得している．

実験は以下の手順で行った．

（1） 被験者に各入力クエリに設定した検索目的を伝える

（2） 被験者が入力クエリの検索結果に含まれる URL 上位 1〜10 位の Web ページデータを閲覧

（3） 被験者が入力クエリの検索結果に含まれる URL 上位 11〜20 の Web ページデータを閲覧

（4） 被験者が入力クエリの検索結果 11〜20 位の各 URL に対し，入力クエリの検索結果 1〜10 位には含まれていない，有用な情報を含んでいるかを判定する

（5） 被験者が代替型検索クエリの検索結果に含まれる URL の上位 1〜10 位の Web ページデータを閲覧

（6） 被験者が代替型検索クエリの検索結果に含まれる各 URL に対し，入力クエリの検索結果 1〜10 位には含まれていない，有用な情報を含んでいるかを判定する

（7） 代替型検索クエリの検索結果上位 1〜10 位に対する判定結果と，入力クエリの検索結果 11 位〜20 位に対する判定結果を比較する

(4)，(6) の判定で得られた結果から，入力クエリの検索結果 11 位〜20 位，代替型検索クエリの検索結果 1〜10 位それぞれにおける適合率を求め，比較を行う．クエリ p の検索結果における適合率は検索結果に含まれる URL10 件に対する，被験者が有用であると判定した URL 数の割合によって求められる．検索結果に含まれる URL10 件のうち，被験者が 3 件を有用と判定した場合，適合率は 0.3 となる．各クエリの検索結果において，被験者が有用な情報を含むと判定した URL 数が多いほど適合率が高くなる．代替型検索クエリの有効性の検証に用いたデータを表 3 に示す．

5. 3. 2　特化型検索クエリの有効性の検証

特化型検索クエリの有効性を検証するため，実験を行った．実験では，入力クエリを用いた検索によって得られる情報と特化型検索クエリを用いた検索によって取得できる情報を比較し，特化型検索クエリを用いた検索によって取得できる情報が入力クエリを用いた検索によって取得できる情報よりも具体化されたか，また入力クエリに関する過去の話題を適切に検索できているかどうかを評価する．実験では以下のデータを 1 セットとして用いる．

- 入力クエリ 1 件
- 入力クエリの特化型検索クエリ 1 件
- 入力クエリの検索結果に含まれる URL の上位 10 件
- 各特化型検索クエリの検索結果に含まれる URL の上位 10 件

実験データとして，上記のデータを 9 セット準備した．各入力クエリの選定は任意で行い，特化型検索クエリは入力クエリを GoogleSuggest に入力することで得られたクエリ群の中から抽出した．実験データとして準備したクエリの検索傾向の時系列データは，GoogleTrends によって取得した．各クエリの検索結果に含まれる URL は，Google を用いて検索を行い取得して

表 4　特化型検索クエリの実験データ

No	入力クエリ	特化型検索クエリ	検索目的
1-1	{秋葉原}	{秋葉原，メイド喫茶}	秋葉原の具体的な過去の話題について知りたい
1-2		{秋葉原，通り魔}	
1-3		{秋葉原，麻生}	
2-1	{オリンピック}	{オリンピック，アテネ}	オリンピックの具体的な過去の話題について知りたい
2-2		{オリンピック，候補地}	
2-3		{オリンピック，北京}	
3-1	{ジブリ}	{ジブリ，ハウル}	ジブリの具体的な過去の話題について知りたい
3-2		{ジブリ，ゲド}	
3-3		{ジブリ，ポニョ}	

表 5　代替型検索クエリの有効性判定実験における入力クエリの検索結果 11〜20 位の適合率

No	入力クエリ	適合率
1	{ディズニーランド}	0.70
2	{サッカー}	0.50
3	{旅行}	0.70
4	{ジブリ}	0.70
5	{オリンピック}	0.60
平均適合率		0.64

いる．

実験は以下の手順で行った．

（1） 被験者に各特化型検索クエリに設定した検索目的を伝える

（2） 被験者が入力クエリの検索結果に含まれる URL の上位 10 件を閲覧

（3） 被験者が特化型検索クエリの検索結果に含まれる URL の上位 10 件を閲覧

（4） 被験者が特化型検索クエリの検索結果に含まれる各 URL に対し，入力クエリの過去の話題についてより具体的な情報を含んでいるかを判定する

（5） 被験者が入力クエリの検索結果と特化型検索クエリの検索結果を比較し，どちらが目的の情報を収集しやすいかを判定する

手順 (4) の判定で得られた結果から，検索結果の適合率を求める．ただし，特化型検索クエリ上位 1〜10 位の検索結果と入力クエリの検索結果上位 1〜10 位に共に含まれる URL が存在する場合，その URL は無効として全体データ数からも除いている．これは，本来特化型検索クエリの検索結果は入力クエリとは関係なく閲覧されることを想定しているためである．特化型検索クエリの有効性の検証に用いたデータを表 4 に示す．

5.3.3　評価と考察

代替型検索クエリの実験結果に対する考察を行う．各入力クエリの検索結果 11〜20 位の適合率の結果を表 5 に，各代替型検索クエリの検索結果 1〜10 位の適合率を表 6 に示す．

代替型検索クエリの検索結果における平均適合率は 0.6 という値となった．この結果から，代替型検索クエリを用いることよって，入力クエリについての話題に関して新たな情報を検索することが可能であると考えられる．しかし，適合率は各代替型検索クエリによってばらつきがあり，特に入力クエリ {サッカー} に関する代替型検索クエリでは全体に値が低くなった．入力クエリ {サッカー} の場合，ワールドカップや日本代表など性質として特化型検索クエリに近いクエリが代替型検索クエリとして抽出されており，ユーザの検索目的とは外れた情報が検索されてしまったため，値が低くなったと考えられる．判定精度の検証においても同様の考察が得られており，このような結果から，現在は代替型検索クエリとして一括りとしている中でも，完全に置き換え可能なクエリ，関連する話題を検索するクエリであるが検索対象が異なるクエリなど，類似の程度や性質を考慮した分類が必要であると考えられる．

表 6　代替型検索クエリの検索結果 1〜10 位の適合率

No	入力クエリ	代替型検索クエリ	適合率
1-1	{ディズニーランド}	{ディズニーリゾート}	0.60
1-2		{ディズニーシー}	0.80
1-3		{東京ディズニーランド}	0.60
2-1	{サッカー}	{サッカー，日本代表}	0.50
2-2		{ワールドカップ}	0.30
2-3		{日本代表}	0.30
3-1	{旅行}	{ツアー}	0.70
3-2		{旅行会社}	0.30
3-3		{京都}	0.50
4-1	{ジブリ}	{宮崎駿}	1.00
4-2		{スタジオジブリ}	0.70
5-1	{オリンピック}	{五輪}	0.80
平均適合率			0.60

表 7　特化型検索クエリの検索結果 1〜10 位の適合率

No	入力クエリ	特化型検索クエリ	適合率	判定値
1-1	{秋葉原}	{秋葉原，メイド喫茶}	0.10	0
1-2		{秋葉原，通り魔}	0.90	1
1-3		{秋葉原，麻生}	1.00	1
2-1	{オリンピック}	{オリンピック，アテネ}	0.50	0
2-2		{オリンピック，候補地}	0.89	0
2-3		{オリンピック，北京}	0.67	1
3-1	{ジブリ}	{ジブリ，ハウル}	0.20	0
3-2		{ジブリ，ゲド}	0.44	1
3-3		{ジブリ，ポニョ}	0.80	1
平均適合率			0.61	

入力クエリの検索結果上位 11〜20 位と代替型検索クエリの検索結果上位 1〜10 位の適合率に大きな差は無かった．これは，検索結果の上位 20 件程度ではあまり URL の内容の重複は発生せず，新しい情報を取得できたためだと考えられる．

次に，特化型検索クエリの実験結果に対する考察を行う．実験で用いたクエリと，各クエリに対する判定結果を表 7 に示す．表中の判定値は入力クエリによる検索結果 1〜10 位，特化型検索検索クエリの検索結果 1〜10 位からそれぞれ得られる情報のどちらが有用であったかについて，被験者が判定した結果を示している．被験者が入力クエリの検索結果の方が有用であると判定した場合は判定値を 0，特化型検索クエリの検索結果の方が有用であると判定した場合は判定値を 1 として表記している．特化型検索クエリによる検索結果上位 1〜10 位の平均適合率は 0.61 という値となった．この結果から，特化型検索クエリを用いることよって，入力クエリについての過去の具体的な話題について検索することが可能であると考えられる．検索結果として入力クエリの検索結果よりも有用だと判定されたの

は 9 クエリ中 5 クエリであった．特に，実験を行った特化型検索クエリの中で，クエリ {秋葉原，メイド喫茶} , {ジブリ，ハウル} の適合率が低い値となった．これは, {秋葉原，メイド喫茶} では実際に存在する店舗のトップページ, {ジブリ，ハウル} では関連商品の紹介ページなどが検索結果に多く含まれ，秋葉原，ジブリについての過去の話題について情報を得ることが出来るページが少なくなったことが原因であると考えられる．この結果から，入力クエリの過去の話題について具体的に検索するために用いられていたクエリを特化型検索クエリとして抽出することは可能であるが，抽出したクエリによって検索される対象が現在までに既に変化していた場合には適切に話題を検索することが出来ないと考えられる．よって，抽出した特化型検索クエリの検索頻度が最も増加していた時期に，特化型検索クエリの検索結果から閲覧されていた URL を検索結果に加えるなど，クエリの検索対象の変化を考慮する必要がある．

また，クエリ {オリンピック，候補地} では適合率が高い値であるにも関わらず，検索結果全体としては入力クエリの検索結果よりも有用性が低いと判定された．これは, {オリンピック，候補地} の検索結果にはオリンピックの候補地選定に関する内容を掲載した URL が多く含まれていたが，1 つ 1 つの URL から得られる情報量が少なく，結果として検索結果全体から得られる情報量が少なくなってしまったためだと考えられる．逆に，クエリ {ジブリ，ゲド} では適合率はそれほど高くない値であるが，検索結果全体としては有用であると判定された．こちらは有用と判定された URL の数は多くないが，それぞれの URL から得られる情報が多く，検索結果全体としては有用であるという判定になったと考えられる．このように，検索結果に含まれる Web ページの数による評価だけでなく，Web ページの質を考慮した評価を行っていく必要がある．

6. おわりに

本稿では，キーワードの時間的類似性を用いてユーザの検索支援を行う手法の提案を行った．キーワード間の時間的類似性は検索エンジンのクエリログからクエリの検索傾向の時系列データを取得し，それを解析することによって判定している．また，時間的な類似のパターンとして検索傾向の全体的な類似，部分的な類似という関係を定義することで，2 つの異なるクエリ間の意味的な関係性を抽出することが可能である．これらの提案手法に関して評価実験を行い，手法の有効性を実際にクエリ用いた検索を行った場合を想定し，検証した．

今後の検討課題として，以下の項目があげられる．

- システムがユーザにクエリを提示するインタフェースの考察
- 提案するクエリ間の時間的関係の応用

まず，本手法ではユーザの検索支援を目的としており，システムが抽出したクエリをどのようなかたちでユーザに対して提示を行うかは非常に重要であると考えられ，実際のインタフェースについて考察する必要がある．また，本手法の応用を考えていく必要がある．本手法で行っている，同一の検索対象に対するクエリの集約と関連の抽出はあるトピックに対する Web 全体での注目度の算出などに応用できると考えられる．あるトピックに対し複数の異なるクエリで検索が行われている場合，トピックの真の注目度を求めるのは難しい．本手法を用いることで，関連を持つクエリを抽出でき，トピックに対する注目度の集約を行うことができる．また，本手法ではクエリの時間的検索傾向の類似性によってクエリ間の関係性を判定しているため，従来の文書データなどの解析からは取得できなかったクエリ間の関係の取得を行うことが可能であると考えられる．本稿では，あくまでクエリを入力したユーザにとって検索に役立つクエリを提示することを目的とした手法の提案を行ったが，時間的検索傾向の類似を用いることで経済，社会，スポーツなどのトピックの分野に捉われず，関連語の取得を行える可能性がある．

謝　辞

本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表します．

文　献

[1] GoogleSuggest: “http://www.google.co.jp/webhp complete 1”.
[2] GoogleSets: “http://labs.google.com/sets”.
[3] Wordnet: “http://wordnet.princeton.edu/”.
[4] 言語工学研究所デジタル類語辞典シソーラス：“http://www.gengokk.co.jp/ruigo.htm”.
[5] D. C. F. Rosie Jones: “Query word deletion prediction”, International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2003), pp. 435–436 (2003).
[6] R. Jones, B. Rey and W. G. Omid Madani: “Generating query substitutions”, Proceedings of the 15th international conference on World Wide Web (WWW2006), pp. 387–396 (2006).
[7] 大島裕明, 小山聡, 田中克己：“Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見”, 情報処理学会論文誌:データベース, Vo.47, No.SIG19(TOD32), pp. 98–112 (2006).
[8] 山口雅史, 大島裕明, 小山聡, 田中克己：“サーチエンジンのクエリログを利用した同位語の発見”, 日本データベース学会 Letters,Vol.5,No.2, pp. 17–20 (2006).
[9] 大島裕明, 山口雅史, 小山聡, 田中克己：“Web 検索とクエリログを併用した同位語発見手法”, 日本データベース学会 Letters, Vol.5,No.4, pp. 37–40 (2007).
[10] N. I. Steve Chien: “Semantic similarity between search engine queries using temporal correlation”, Proceedings of the 14th international conference on World Wide Web (WWW2005), pp. 2–11 (2005).
[11] C. Z. Xuanhui Wang: “Learn from web search logs to organize search results”, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007), pp. 87–94 (2007).
[12] P. H. Kenneth Ward Church: “Word association norms, mutual information, and lexicography”, Proceedings of the 27th annual meeting on Association for Computational Linguistics, pp. 76–83 (1989).
[13] S. C. H. H. Qiankun Zhao: “Time-dependent semantic similarity measure of queries using historical click-through data”, Proceedings of the 15th international conference on World Wide Web (WWW2006), pp. 543–552 (2006).
[14] 甲谷優, 湯本高行, 小山聡, 田中克己：“クエリログを用いた web ページの典型的クエリの抽出とその応用”, 電子情報通信学会：

第 19 回データ工学ワークショップ (DEWS 2008) (2007).
[15] 甲谷優, 大島裕明, 小山聡, 田中克己：“典型的クエリを用いた web ページの重要箇所特定”, 情報処理学会データベースと Web 情報システムに関するシンポジウム (DBWeb 2007) (2008).