Vol.2010-DBS-151 No.1
2010/11/12

レアクエリを対象とした非クリック分析によるクエリ推薦方式の提案

宮原伸二[†]　片渕典史[†]　片岡良治[†]

近年，検索ログを用いたクエリ推薦に関するサービスが広く普及し，ユーザは推薦されるクエリを用いて効率的に検索できるようになった．しかし，これまで実現されたクエリ推薦手法は，検索ログ内で利用頻度の高いクエリに対して有効であり，利用頻度の少ないレアクエリには適用が困難である．そこで本研究では，レアクエリを対象としたクエリ推薦手法として，クリックグラフとスキップグラフによる特徴量を用いたクエリ推薦手法を提案する．また，提案手法の有効性を確認するため，検索ログを用いた評価実験を行う．

A Method that Suggests the Related Queries for a Rare Query Using Skipped URL Information

Shinji Miyahara[†]　Norifumi Katafuchi[†]　Ryoji Kataoka[†]

Recently, the query suggestion service is widely widespread on the Internet. The users can retrieve web pages efficiently by using a suggested query. However, while the conventional methods of the query suggestion are effective for a query with high click frequency in the retrieval log, it is difficult to apply these methods for a query with low click frequency. Therefore we propose the query suggestion method using the click graph and the skip graph. Evaluations are also conducted using a commercial search engine log.

1. はじめに

近年，検索ポータルサイトにおける検索ログやクリックログを利用し，ユーザの検索支援に役立てる取り組みが注目されている．特に，ユーザの検索活動を効率化するための支援として，ユーザの入力するクエリに対し，他のユーザが頻繁に利用するクエリを推薦したり，絞り込みのクエリを推薦する試みが広く行われている．この推薦によって，ユーザは欲しいWebページに対する効果的なクエリを得ることができる．それにより，欲しいWebページに対するコンテキストが少ないユーザでも，効率的に検索活動が行える．

一般に，検索ログを用いたクエリ推薦では，検索ログにおけるURLのクリック数であるCF(Click Frequency)やクエリの利用頻度であるQF(Query Frequency)を用いて，クエリとURL間の結びつきを確率的に求める研究が行われている2)5)．このクエリとURL間の確率により，頻繁にクリックされるURL集合にアクセスするための効果的なクエリを抽出できる．また，URLを介したクエリ同士における到着時間を用いたクエリ推薦にHitting Timeを利用した手法がある3)．この手法では，クエリ間のHitting Timeを再帰的に計算し，得られたHitting Timeの小さいクエリ同士に強い関連があると考えている．この手法では，入力したクエリで得られるURL集合と関連が強いURL集合に結び付くクエリが推薦される．

これら従来のクエリ推薦手法では，主にpopular query（頻繁に利用されるクエリ）を対象としており，このpopular queryに対しては効果的に機能している．その理由としては，推薦するクエリを選出する際のグラフ分析において，クエリとURLの結びつきやCFなどの情報が豊富に存在するため，確率的なアプローチが有効に機能していると言える．それに対し，rare query（利用頻度の少ないクエリ）に対しては，クエリとURLの結びつきが少なく，CFの値が小さいためクエリを推薦するための分析が困難である．また，rare queryを利用するユーザも少ないため，一部のユーザの利用状況や特定の用途に大きく影響を受ける問題がある．このrare queryを対象としたクエリ推薦の先行研究としてYangらのスキップグラフを用いた手法がある1)．この手法では，ユーザに検索結果として提示したURLに対し，ユーザがクリックしていないURLで構成したスキップグラフと一般のクリックグラフの特徴量を結合させてrare queryに対する問題を解決している．しかし，この手法ではユーザに提示したクリックしていないURLのログが必要であり，一般に利用されるクエリとクリックしたURLの検索ログには適用できない．

そこで本研究では，ユーザが投入したクエリとクリックしたURLで構成される検索ログを用いたrare queryに対するクエリ推薦手法を提案する．提案手法では，クリック関係のないクエリとURLに対し検索順位に基づいた重みでエッジを作成するスキップグラフを用いる．また，各URLに対し，ユーザがクリックした時間が新しいほど重要なURLとし，スキップグラフのエッジに対してクリックした時間に応じた重みを加える．これらの検索順位とクリックした順位を考慮

[†] 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

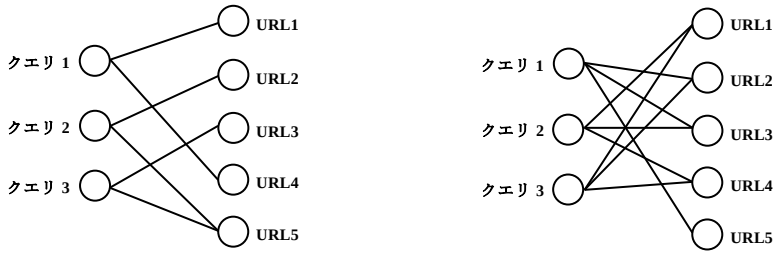
したスキップグラフによる特徴量とクリックグラフによる特徴量を結合してクエリを推薦する. また, 提案手法を商用の検索ログに適用し, 人手評価データを用いた評価実験において提案手法の有効性を実証するとともに, 考察を行う.

2. 関連研究

検索ログを用いたクエリ推薦の研究は数多く存在し，大きく popular query を対象とした研究と rare query を対象とした研究に分けられる．

popular query を対象とした代表的な研究として，Deng らのクリックエントロピを利用した研究がある 2)．この研究では，クエリの利用回数と URL をクリックした回数を基にクエリと URL 間のクリックエントロピを算出し,クリックグラフでのエッジの重みに利用している．クリックエントロピは，扱う情報の規模が大きいほど有効に機能するため，この手法は popular query へのクエリ推薦に有効に機能している．また，ユーザが投入したクエリに対し，推薦するクエリの多様性を目指した研究に Mei らの Hitting Time を用いた手法がある 3)．この手法では，検索ログ内においてクエリとクリックされた URL で構成されたクリックグラフにおいて，Hitting Time と呼ばれる到着時間を用いて，到着時間の小さいクエリ同士が強く関連すると解釈している．この研究は，ユーザが投入したクエリに対し，多様性をもったクエリを推薦することでユーザの検索範囲の拡大に貢献している．しかし，この手法も Deng らの研究と同様，popular query に対して有効であり，rare query へは不向きである．

rare query を対象とした代表としたクエリ推薦の研究として，Yang らのスキップグラフを利用した手法が存在する 1)．この手法では，rare query に関連するクエリが少ない問題に着目し，関連するクエリを多くするためユーザに提示したクエリの検索結果上位の URL は関連する URL とし, クエリとユーザがクリックしていない URL で構成したスキップグラフを作成している．



(a)クリックグラフ　　(b)スキップグラフ

図 1　クリックグラフとスキップグラフ

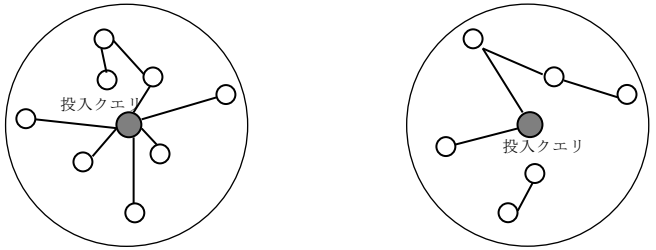
図 1 にクリックグラフとスキップグラフの例を示す．図 1 の（a）は，検索ログ内でクリック関係のあるクエリと URL のグラフである．図 1 の（b）は，クリックグラフグラフにおいてリンク関係のないクエリと URL を連結させたグラフである. このスキップグラフとクリックグラフの特徴量を用いて, rare query に対するクエリ推薦を実現し, 評価実験による有効性を確認している．この手法では，ユーザが投入したクエリとクリックした URL だけでなくユーザに提示した検索結果の URL が必要なため，手法を実現する環境に大きな制限を受ける．

3. rare query のグラフ特徴

検索ログを用いたクエリ間のグラフは, popular query と rare query では特徴が大きく異なる．この 2 種類のクエリでは，特にクリック回数やクリックされる URL の数，クエリ間に存在するリンク関係の特徴が異なるためクエリ推薦でのアプローチが異なる．

popular query では，クエリの利用頻度や URL のクリック回数が多いため，ランダムウォークなどの確率的なアプローチが有効に機能する 4)．それに対し rare query は利用頻度が少なく，クリックされた URL の数も少ない．そのため，クエリと URL で構成したクリックグラフを用いて，ランダムウォークなどによる確率的なアプローチを用いてクエリ間の関連度を算出しても精度が低い問題がある．

Vol.2010-DBS-151 No.1
2010/11/12



(a)popular query のリンク関係図　　(b)rare query のリンク関係図

図 2　2 種類のクエリのリンク関係図

popular query と rare query における，関連するクエリとのリンク関係について図 2 を用いて説明する．図 2 の中心に位置する投入クエリはユーザの投入したクエリを表しており，周囲に複数の円で表示したクエリは人手評価などで評価した関連の近さを距離で表現している．また，クエリ間のリンクは，検索ログ内で URL を介したクエリ間の結びつきを表している．図 2 の popular query のリンク関係図では，URL を介して結びつくクエリが多く，ユーザが投入したクエリと関連の強いクエリへのリンクが多く存在する．それとは逆に，rare query では，関連するクエリとのリンクが非常に少ない．その理由としては，rare query ではクリックされる URL が少ないため結びつくクエリが非常に少ないと考えられる. そのため, rare query を対象としたクエリ推薦では，クリック関係のないクエリと URL の結びつきを考慮することが必要となってくる．

4. 提案手法

4.1 提案手法の概要

3 章の rare query の特徴から，rare query に対し検索ログを用いてクエリを推薦するには, クエリとクリックされた URL のみを用いたクリックグラフによる分析では困難である. そのため, URL を介してリンクしないクエリ同士をリンクさせる必要がある. そこで我々は，クエリとクリックされた URL で構成されるクリックグラフと，クエリとクリックされていない URL で構成されるスキップグラフを用いたクエリ推薦手法を提案する．また，これらクリックグラフとスキップグラフの特徴量を合成したものを rare query のクエリ推薦に用いる．

4.2 クリックグラフ

クリックグラフは，rare query においても重要な要素である．ユーザが投入したクエリとクリックした URL は，ユーザの明示的な振る舞いであり，クリックグラフから得られるクエリ間の特徴量は rare query においても一般性をもったクエリ間の特徴量と言える．クリックグラフの作成方法について説明する．クリックグラフは，検索ログにおいてユーザが投入したクエリと，その検索結果に対してユーザがクリックした URL で構成されるクリックグラフである．このクリックグラフにおけるエッジは，ユーザがクエリを投入してクリックしたことを示している．また，各エッジには重みp_{ij}を付加しており, クエリの投入回数と URL へのクリック回数を基に下記の式で算出される．

$$p_{ij} = \frac{c_{ij}}{d_i}$$

上記の式において, c_{ij} はクエリ i から URL j へのクリック回数であり, d_i はクエリ i を投入して URL をクリックした総数を示している. そのため, p_{ij} はクエリ i から URL j への遷移確率を意味している．上記の式で求められる遷移確率を用いて URL を介したクエリ間の遷移確率を下記の式で求める．

$$p_{ij} = \sum p_{ik} \times p_{kj}$$

上記の式の k は，クエリ i とクエリ j を結びつける URL k を表しており，検索ログ内ではクエリ i とクエリ j の両クエリで URL k がクリックされている．

4.3 スキップグラフ

スキップグラフは, 検索ログ内ではクリック関係のないクエリと URL 間にエッジを作成したものであり，エッジの作成方法が重要となる．我々は，ユーザに提示したクエリの検索結果において, 検索結果上位の URL はクリックグラフ内の他のクエリにおいても重要な URL と考え，クリック関係にないクエリと URL 間に検索順位を基にしたエッジを作成する．ここで，クリック関係にないクエリと URL にエッジを作成した場合，多くのエッジが作成されるため，グラフ全体の特徴が損なわれる問題も発生する．そこで，スキップグラフに URL をクリックした時間の新しさに基づいてエッジの重みを加えることで，作成したエッジによるグラフの特徴量の低下を防ぐ．

スキップグラフの作成方法について説明する．スキップグラフで作成するエッジには検索ログ内での URL の検索順位とクリックされた時間に応じた重みを付与する. 検

索順位に応じたエッジの重みは，下記の式で算出する．

$$w_{ij} = \log\frac{1}{rank} \times \log\frac{1}{URL_{time}}$$

上記の式での第一項は，クエリの検索結果に対する URL の検索順位である *rank* によって決定され，この検索順位が上位であるほど値が大きくなり重要と考える．また，第二項は，URL が最も最近クリックされた時間である *URLtime* で決定し，この時間が時間的に新しいほど重要度が高いと考える．

このエッジの重みをクリック関係のないすべてのクエリと URL に対して作成する．また，この重みはすべてのエッジに対して算出後，0 から 1.0 までの値に正規化し，この値をクエリから URL への遷移確率として用いる．

4.4 特徴量の合成

スキップグラフ特徴量は，クエリのポジティブな特徴量であり，クリックグラフの特徴量と同様に扱えるため，スキップグラフとクリックグラフの特徴量を線形和で合成して，rare query のクエリ推薦の特徴量として用いる．この合成では下記の式で合成する．

$$R = \alpha R_{click} + (1 - \alpha) R_{skip}$$

上記の R_{click} はクリックグラフによる特徴量で表される行列であり，各クエリ間の遷移確率で構成される．また，R_{skip} はスキップグラフによる特徴量で表される行列であり，URL の検索順位とクリック時間から算出されたクエリ間の遷移確率である．この 2 つのグラフの遷移確率をαを用いて線形結合する．この遷移行列を用いてクエリ *i* への推薦クエリを抽出する際は，行列内のp_{ij}で最も大きな値をもつ *j* におけるクエリ *j* を推薦クエリとする．

5. 評価実験

本研究では，検索ログを用いてベースライン手法と提案手法を適用し，クエリの推薦結果に対して人手評価データを用いた評価実験を行った．検索ログは商用の検索サービスログを用いた．

ベースライン手法はクリックグラフを用いた最も基本的な手法である CF・IQF を用いた．また，ベースライン手法，提案手法の各手法において，評価クエリに対する検索ログ内でクリック関係にある URL から，それら URL とクリック関係のあるクエリの検索ログを抽出し，それら抽出した検索ログを用いてクリックグラフ，スキップグラフをそれぞれ作成した．

5.1 データセット

検索ログは商用の検索サービスでの 1 ヵ月間のログを用いた．評価クエリは，1 ヶ月間の検索ログからクエリの投入回数が 50 回以下のクエリからランダムに抽出した 150 クエリを評価クエリとして利用した．

5.2 人手評価データ

人手評価データは，150 件の評価クエリに対し，各クエリに対する各手法での推薦クエリ上位 5 件との関連度を人手で評価したデータである。推薦クエリとの関連度は、評価クエリに対し，各推薦クエリの検索結果 150 件と評価クエリとの関連度を 3 人の評価者で 4 段階評価し，その平均値を評価値として用いた．

5.3 実験結果

表 1　各手法における推薦クエリ上位 5 件の評価結果

	平均評価値	NDCG@5
ベースライン手法	2.59	0.588
提案手法	2.61	0.590

表 1 に実験結果を示す．表 1 では，ベースライン手法，提案手法で推薦した上位 5 件の推薦クエリに対する平均評価値，NDCG@5 における評価結果を示す．

推薦クエリの平均評価値では、提案手法がベースライン手法の結果よりも高い精度を示しているが，有意性を示すまでに至らなかった．

表 1 の NDCG@5 からは、提案手法がベースライン手法の結果よりも高い精度を示しているが，平均評価値と同様に有意性を確認することはできなかった。

6. 考察

人手評価を用いた評価実験では，提案手法の有効性は確認できなかった．そこで，評価実験で用いた検索ログの特徴を基に実験結果を分析し提案手法の考察を行う．

表 2　検索順位に応じた評価結果

	rank 平均 10 未満	rank 平均 10 以上
ベースライン手法	0.610	0.569
提案手法	0.578	0.631

表 3　クエリの推薦例

推薦順位	ベースライン手法	提案手法
1	ヒストリア	歴史秘話ヒストリア
2	蟹工船　nhk	nhk ヒストリア
3	歴史秘話ヒストリア	歴史秘話
4	明治人々	nhk ヒストリー
5	nhk ヒストリー	ヒストリア

まず始めに，各評価クエリに対するクリックグラフ内の URL に対し，平均検索順位に着目したベースライン手法と提案手法の評価結果について考察する．表 2 に，平均検索順位が 10 位未満の評価クエリと，平均検索順位 10 以上の評価クエリに対する各手法の評価結果を示す．平均検索順位が 10 未満の評価クエリに関しては，ベースライン手法と比較し提案手法の結果が大きく下回っていた．それとは逆に平均検索順位が 10 以上の評価クエリに対しては，提案手法が有効に機能した．提案手法は，スキップグラフにおいて平均検索順位を基にエッジの重みを算出していため，検索順位の偏りの少ないクリックグラフでは，スキップグラフの特徴量が有効に機能していないためだと考えられる．

次に，提案手法におけるクリックグラフの特徴量とスキップグラフの特徴量の結合に利用する α の値について考察する．Yang らの手法では，この α に 0.8 程度の値に設定しており，クリックグラフの特徴量を大きく採用している．それに対し，提案手法では，評価実験での α の値は 0.4 に設定した．この設定は，クリックグラフよりもスキップグラフの特徴量を大きく採用することを示しており，クリック関係のないクエリと URL にも関連性があると考えられる．この結果から，提案手法は rare query の利用頻度や関連するクリック関係にあるクエリと URL の数が少なくても適用可能であり，より広範囲の rare query に対して有効に機能すると考えられる．

また，各評価クエリに対する各手法で推薦されるクエリについて考察する．ベースライン手法では，多くの推薦クエリにおいて人手評価値の低いクエリが推薦されることが確認できた．この評価値が低いクエリが推薦される例として，表 3 にテレビ番組を表したクエリ「歴史ヒストリア」に対する各手法の推薦結果を示す．表 3 では，提案手法において，テレビ番組の正式名称である「歴史秘話ヒストリア」が推薦されていることに対し，ベースライン手法では，この番組で取り上げられた内容の「蟹工船 nhk」や「明治人々」が推薦されている．この理由は，検索ログにおいて一部のユーザが番組で取り上げられた特定の内容に関する URL を過大にクリックしていたため，クリックグラフにおけるクエリの特徴量が大きくなったためだと考えられる．それに対し，提案手法では番組の一部の内容を示すクエリでなく，正しい番組名を示すクエリが推薦されることを確認した．この事実は，一部のユーザによるクリックの影響が，スキップグラフによる特徴量で低減されたためと考えられる．

7. おわりに

本研究では，検索ログにおいて利用頻度の少ない rare query を対象としたクエリ推薦手法を提案した．提案手法では，検索ログ内で rare query と関連するクエリとのリンク関係がないことに着目し，クリック関係のない URL へのエッジを作成するスキップグラフを用いたクエリ推薦手法を実現した．この手法では，クリック関係のない URL でも，他のクエリとのクリック関係において検索順位が上位であれば重要な URL として捉え，検索順位に応じたエッジの重みを作成して関連するクエリの到達機会を作り出した．また，クリック関係のない URL へのエッジを作成した場合に発生するエッジ数の増加で確率的なアプローチが有効に機能しない問題に対し，URL をクリックした時間が新しいほど重要な URL とし，エッジの新たな重みとして利用した．

評価実験では，商用の検索ログを用いてベースライン手法と提案手法を比較した．実験結果からは，提案手法の有効性は確認できなかったが，クリックされた URL の検索順位の平均値が大きいクエリに対しては有効性が確認された．また，提案手法で推薦するクエリは，一部のユーザの過大クリックによる影響を受けにくいことを確認しており，rare query におけるクエリ推薦がロバストであることが示された．

今後は，スキップグラフの精度向上を目指して，クリック関係のない URL の重要度のモデル化を行い，rare query に対するクエリ推薦精度のさらなる向上を目指す．さらに，より多くの評価データを用いた大規模な評価実験を行う．

参考文献

1) Yang Song, Li-wei He, Optimal Rare Query Suggestion With Implicit User Feedback, WWW, April, ACM, 2010
2) Hongbo Deng, Irwin King, Michael R. Lyu, Entropy-baiased Models for Query

Vol.2010-DBS-151 No.1
2010/11/12

Representation on the Click Graph, In Proceedings of SIGIR'09, pages 339-346, ACM, 2009.
3) Qiaozhu Mei, Dngyong Zhou, Kenneth Church: Query suggestion using hitting time, In Proceedings of CIKM '08, pages 469-478, ACM, 2008.
4) N. Craswell and M. Szummer, Random walks on the click graph, SIGIR, pages 239-246, 2007
5) S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, journal of Documentation, 60:503-520, 2004