5B-5

曖昧性を考慮した多様なクエリ推薦

Diversified Query Suggestion considering Query Ambiguity

江田 毅晴 † Takeharu Eda
俵本 一輝 ‡ Kazuki Tawaramoto
宮原 伸二 † Shinji Miyahara
片渕 典史 † Norifumi Katafuchi
片岡 良治 † Ryoji Kataoka

1 はじめに

ユーザの検索活動を支援する方法としてクエリ推薦 (Query Suggestion/Recommendation) がある．ユーザが入力したクエリの次に入力すべきクエリを推薦し，キーボード入力の手間を省き，検索行動を効率化する．しかしながら，ユーザの情報要求は多様であるとともに曖昧性を持つため，入力されたクエリのみから情報要求を特定し，適切なクエリを推薦することは困難である．

既存手法は，主にクエリと URL のつながりから構成される 2 部グラフのランダムウォークによるものが主流である [3, 4, 5]．例えば，RWR (Random Walk with Restart) では，入力されたクエリを中心としたクラスタ (エゴイスティッククラスタ) への帰属確率を求めることによりクエリを推薦するが，ユーザの情報要求はモデル化されておらず，単純にグラフの連結性に基づいたクエリが推薦されることになる．結果として，似たクエリばかり推薦されるケースがある．

そこで，本研究では，トピックモデルを用いた確率的クラスタリングによりユーザの情報要求を抽出する．入力クエリのクラスタの生起確率分布によりクエリの曖昧性を判定し，クエリ推薦を行う．

2 提案手法

提案手法は，下記の 3 つの仮説に基づく．

1. 検索セッションで情報要求が満たされる．
   ユーザが検索を開始し終了するまでの一連の検索セッションで何らかの情報要求が満たされると考えるのは自然である．この仮説に基づいて，我々はまずクリックログから検索セッションを切り出す．
2. 検索クエリのみで情報要求を一意に特定することはできない．
   「京都」というクエリが，京都の観光情報に関する要求なのか，それとも京都への行き方に関する要求なのか，クエリそのものから一意に特定することはできない．すなわち「京都」というクエリは複数の情報要求を持つ可能性がある．この仮説により，我々はトピックモデルを用いてクエリを確率的にクラスタリングすることを考えた．
3. クエリが曖昧なときは，複数の情報要求を提示し選択肢を増やすことが有効である．
   入力されたクエリが高い確率で 1 つの情報要求を持つと判断される場合，その要求を満たすことができるクエリにユーザは満足すると考えられる．一方で，複数の情報要求を持つ可能性がある場合には，それぞれの情報要求を満たすクエリ集合から代表的なクエリを推薦しユーザの選択肢を増やすことでユーザの満足度が向上する．

提案手法の概要を図 1 に示す．クリックログからセッションを切り出し，セッション内のクエリと URL からなる 2 部グラフを構築する．このグラフに対して，トピックモデルにより確率的クラスタを構築する．本研究ではこのクラスタをユーザの情報要求であるとみなす．実際の推薦処理は，入力クエリからクラスタへの生起確率分布に基づいて行う．入力クエリから特定のクラスタが強く生起する場合は入力クエリの情報要求が特定されたと考え，そのクラスタからクエリを推薦し，複数のクラスタが生起する場合には入力クエリは曖昧であると考え，複数の情報要求からラウンドロビンでクエリを推薦する．

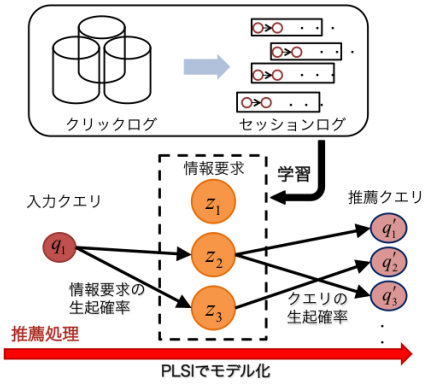

図 1 提案手法の概要．

2.1 クリックログからのセッショングラフ構築

関連研究を踏まえ [4]，同一セッション ID のクリックログでクエリ投入間隔が 10 分以内となるものを検索セッションとした．単一の検索セッションにはノイズが含まれることもあるため，多数の検索セッションをマージしたセッショングラフを構築する．今回は，「入力として与えた 1 つのクエリを含む」という条件で検索セッションを切り出す．セッショングラフには入力クエリが持つ可能性のある複数の情報要求を満たすクエリが含まれていると想定する．

2.2 セッショングラフへのトピックモデルの適用

セッショングラフに含まれる複数の情報要求を，トピックモデル (topic model) [2] により確率的に浮かび上がらせる．今回は，トピックモデルとして PLSI [1] を用いた．PLSI は，隠れ変数による共起のモデル化であり，クエリを $q_i \in Q$, URL を $r_j \in R$ とすると，下記の式により表される．

$$p(q_i, r_j) = \sum_{k=1}^{n_Z} p(z_k)p(q_i|z_k)p(r_j|z_k)$$

ただし，$z_k (1 \leq k \leq n_Z)$ は隠れ変数であり，確率的クラスタすなわち情報要求を表すと考える．隠れ変数を含む確率は EM アルゴリズムを用いて推定する [1]．これにより，クエリ q_i からクラスタ z_k が起きる確率 $p(z_k|q_i)$ は，ベイズ則を用いて下記のように求められる．

$$p(z_k|q_i) = \frac{p(q_i|z_k)}{\sum_{l=1}^{|Q|} p(q_l|z_k)}$$

$p(z_k|q_i) (1 \leq k \leq n_Z)$ は確率分布になっており，クエリ q_i からそれぞれの情報要求が起きる確率を表したものであ

† 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
‡ 京都大学大学院 情報学研究科 社会情報学専攻

る．クラスタ数 n_Z は，今回は，数回試行した結果を目視評価により経験的に定めた．クラスタ数の自動的な決定方法としては，例えば，AIC を用いる方法が考えられる．

2.3 クエリ推薦アルゴリズム

基本的には，入力クエリ q_i から起きる確率 $p(z_k|q_i)$ が高い情報要求 z_k を選択し，z_k から代表的なクエリ q_m を推薦する．代表的なクエリの選択方法としては，情報要求からクエリが起きる確率 $p(q_m|z_k)$ をそのまま用いる方法やセッショングラフに RWR を適用した定常確率を用いる方法が考えられる．今回は，$p(q_m|z_k)$ が高いクエリを代表クエリとして利用する．

$p(z_k|q_i)$ が特定の要求に高い確率を持たず，複数の要求に属する可能性がある場合には，複数の要求からラウンドロビンで代表的なクエリを推薦する．これにより，推薦クエリの多様性が高まり，ユーザの選択肢が増えるため，情報要求を明確に表現するクエリの作成を支援できる．推薦アルゴリズムを下記に示す．

---

input: Sessing graph: SG, query: q,
Num. of recommended query: n_q ,
Threshold: $p_{th}(0 < p_{th} \leq 1)$
output: Suggested query list: $L(|L| = k)$
1 $p(z|q) > p_{th}$ となるクラスタ集合を取得し， $p(z|q)$ の降順に，リスト CZ に挿入する．
2 $i = 0$ (取得済みクエリ数)
3 **while** ($|CZ| > 0 \& i < n_q$)
4 $j := i$ を $|CZ|$ で割った余り
5 $k := i$ を $|CZ|$ で割った商
6 $CZ[j]$ から $k + 1$ 番目に $SCORE$ が高いクエリを L に追加.
7 i = i+1
8 **endwhile**
9 L を出力する．

アルゴリズムでは，CZ に該当するクラスタが一つの場合には，クエリに曖昧性が少なく情報要求が特定されたと判断し，そのクラスタのみから $SCORE$ の高い上位 n_q 件のクエリが推薦される．CZ に複数のクラスタが含まれる場合には，クラスタ集合が入力クエリから生起しやすい順にソートされ，ラウンドロビンで順番に一つずつクエリを選択し出力する．$|CZ| = 0$ となるケースもあり，その場合は入力クエリは曖昧すぎるため情報要求は特定されないと考え，出力を空とする．CZ を特定する際のしきい値 p_{th} は，推薦クエリの多様性に関するパラメータになり，小さくするほど複数のクラスタが選択されるためより多様なクエリが推薦されることになる．各クラスタ内で $SCORE$ が高い順に事前にクエリをソートしておくことができるので，推薦処理自体は非常に高速に実行することができる．

3 実験

クラスタの精度, 推薦精度, およびクエリの曖昧性について実験を行った．データセットとして，商用の検索エンジンのクエリログから，一ヶ月の期間内で投入頻度の高かったクエリを利用してセッショングラフを構築した．

3.1 クラスタの精度

表 1 にトピックモデルによって抽出された確率的クラスタの代表クエリを示す．意味的なまとまりがあるクラスタが見られる．

3.2 推薦精度

比較対象として、RWR を用いた．今回は，セッショングラフ構築に用いたクエリをテストクエリとし, そのクエリを含む 100 セッションをランダムに選び，同一セッション内クエリへの適合率を調べた．RWR も同一のセッショングラフに対して適用した．図 2 に適合率の結果を示す．

表 1 ”サッカー” を含むセッショングラフから構築した確率的クラスタ

0	“サッカー 世界ランキング”, “アルゼンチン サッカー”
1	“コラム サッカー”, “2ch サッカー” “サッカー ワールドカップ 日程”
2	“サッカー スポーツナビ” “サッカー 動画”, “日本コカコーラ”
3	“サッカー 日本代表”, “サッカー 本田”
4	“サッカー スパイク”, “田中理恵”
5	“2 チャンネル”, “サッカー ヤフースポーツ”
6	“yahoo youtube”
7	“サッカー ワールドカップ”, “ワールドカップ”
8	“ワールドカップサッカー” “fifa”, “サッカー パラグアイ”
9	“サッカー 関西大学”, “藤田咲”

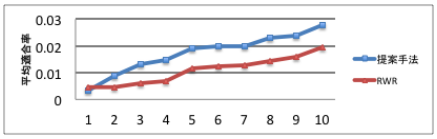

図 2 上位 k 件での平均適合率.

3.3 クエリの曖昧性

入力クエリと提案手法により選択された確率的クラスタ数 (=$|CZ|$) および代表クエリの一例を表 2 に示す．クラスタ数が多く選択されるほど，入力クエリが曖昧と判定され, より多様なクエリが推薦される．一方で，クラスタ選択パラメータ p_{th} の設定により，推薦クエリの多様性をチューニングすることもできる．

表 2 選択されたクラスタ数 $|CZ|$ と推薦クエリの例.

入力クエリ	p_{th}	$\|CZ\|$	推薦クエリ
“ワールドカップ”	0.1	4	“サッカー ワールドカップ”, “ワールドカップ 日程”, “W 杯 ワールドカップ”, “fifa”
“素直になれなくて”	0.1	3	“ドラマ 素直になれなくて”, “mother”, “主題歌 素直になれなくて”, “月の恋人”
“akb48”	0.25	1	“akb48 総選挙”, “akb48 選挙”, “akb48 総選挙”, “akb48 中間発表”
“akb48”	0.2	2	“akb48 総選挙”, “akb48 選挙速報”, “akb48 選挙”, “akb48 人気投票”
“akb48”	0.15	4	“akb48 総選挙”, “akb48 選挙速報”, “akb48 選抜総選挙”, “akb48 メンバー 一覧”

4 まとめ

本稿では，トピックモデルを用いてユーザの情報要求を抽出し，入力クエリの曖昧性を考慮したクエリを推薦する手法を示した．提案手法では，クエリログから作成したセッションログから構築された情報要求クラスタへの入力クエリからの生起確率分布によりクエリの曖昧性を判定する．これにより，クエリから情報要求が特定される場合にはその情報要求を満たすクエリを，曖昧な場合には多様なクエリを推薦することができる．

参考文献

[1] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.
[2] Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. *Handbook of Latent Semantic Analysis*. Psychology Press, 2007.
[3] Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. Query suggestion using hitting time. In *Proc. CIKM*, 2008.
[4] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. Clustering query refinements by user intent. In *Proc. WWW*, 2010.
[5] 今井, 戸田, 関口, 望月, 鈴木, 今井. Web 検索サービスにおける多義的なクエリ推薦手法. In *Proc. DEIM*, 2010.