複数意図のエンティティクエリに対する絞り込み検索のためのクエリ生成法の提案

豊田 樹生　齋藤 純　小松 広弥　熊谷 賢　菅原 晃平
ヤフー株式会社
{itoyota, junsait, hkomatsu, kenkumag, ksugawar}@yahoo-corp.jp

概要

本論文では複数意図を持つエンティティクエリに対する絞り込み検索のためのクエリ生成に取り組む．複数の正例生成器とラベル未付与事例生成器を組み合わせることで訓練事例を自動生成することを提案する．クエリの組の CRR(Cumulative Reciprocal Rank) の差を含めた複数の素性を用いて Random Forest[4] による PU(Positive Unlabeled) 学習を行う．提案手法により CRR の差を単独で用いる場合よりも F 値が 4.4 ポイント向上することを示す．

1　はじめに

ウェブ検索においては，しばしば複数意図を持つエンティティクエリ [11] が発行される．例えば，メディア作品名のクエリ “ゆるキャン△” では，漫画・ドラマ・アニメなどの複数の意図がある．人物名のクエリ “森麻季” では，アナウンサーや歌手などの複数の意図がある．このようなクエリに対して，それぞれの意図に対応した絞り込み検索をできるようすることは検索体験を向上させるうえで重要である．先行研究の複数意図クエリのブレンド検索 [12] や関連クエリの自動補完 [10] はエンティティクエリを考慮したモデリング・評価が十分であるとは言い難い．そこで，本論文では次のような貢献を行う：

(i)複数意図のエンティティクエリに対して絞り込み検索を行えるようにするための再検索クエリの生成方法として，検索ログに蓄積された元クエリ・再検索クエリの組を順位付けして利用することを提案する．

(ii)複数のラベル生成器を用いた訓練事例の自動生成法を提案し，PU(Positive Unlabeled) 学習を行えることを示す．

(iii)Random Forest[4] により学習を行い，単独のラベル生成器を用いた場合と比較して F 値が 4.4 ポイント向上したことを報告する．

2　問題設定

元クエリ q_a に対して，再検索クエリ q_b の順位付けされたリストを生成する．

このとき，次のクエリの要件を全て満たす候補のみを選択する：1) q_a，q_b はいずれもエンティティクエリである 2) q_a は複数のエンティティを指している 3) q_b は特定のエンティティを指す意図の絞り込みのクエリである

3　挑戦的課題

本タスクの挑戦的課題を次に示す：

意図の絞り込みではない再検索の除外　1) 力点の変化を判定できなければならない．例えば “アガサクリスティ”→“アガサクリスティ ねじれた家” の場合は付加された “ねじれた家” に力点が変化している．こういった組は除外するべきである．2) 周辺語を含まない部分一致はクエリの表層だけでは判定が難しい．例えば “東京”→“東京タワー” の場合は意図を絞り込んでいるわけではないため除外するべきである．一方，同じ部分一致でも取り違える可能性の高い組は絞り込みのための再検索として残す必要がある．例えば “ディーゼル”→“ディーゼルエンジン” などが挙げられる．

知識外の再検索候補の順位付け　所与の知識ベースに格納されているエンティティが再検索先の候補として最もふさわしくなるとは限らない．例えば “RHP” というクエリでは “バイトル RHP”(ホームページ作成サービス) などが再検索クエリの候補として挙げられる．しかし，これと対応するエンティティは Wikipedia などの知識ベースには格納されていない．

4　フレームワーク

図 1 に提案手法のフローチャートを示す．手順は次のとおりである．まず，検索ログを二種類取得する．一つ目は素性抽出用のログである．あらかじめ

学習時に参照するための素性を保存しておく．二つ目は順位付け対象となる元クエリおよび再検索クエリの組を取得するためのログである (4.1 節)．このログから取得された順位付け対象のクエリの組に対し，複数のラベル生成器を用いて正例とラベル未付与の事例とに分割する (4.2 節，4.3 節)．さいごに，PU 学習 [4] を行い，順位付けのための回帰器の学習およびそれを用いた予測確率の付与を行う (4.4 節)．

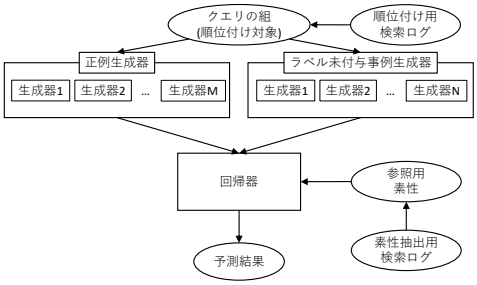

図 1: フローチャート

4.1 検索ログの取得

ウェブ検索のセッションログ[1)]を取得する．このログから再検索クエリ q_b の発行された時刻 $t(q_b)$ と元クエリ q_a の発行された時刻 $t(q_a)$ の差が 30 秒以内のもののみを抽出する．

4.2 正例生成器

次の 3 つの正例生成器を提案する：

4.2.1 元クエリに対するエンティティリンカー

元クエリを内製のエンティティリンカー [14, 15, 16] の入力とし，エンティティ ID を出力する．知識ベースからエンティティ ID と紐づく正式名称を取得する．元クエリが正式名称に対する部分一致文字列になっている場合は元クエリ，正式名称の組を正例とする．

4.2.2 クエリの組に対するエンティティリンカー

元クエリと再検索クエリの組を内製のエンティティリンカー [14, 15, 16] の入力とし，それぞれのエンティティ ID を取得する．元クエリと再検索クエリでそれぞれ異なるエンティティ ID を出力している組を残す[2)]．さいごに，次の条件をすべて満たす組を正例とする：1) 人物エンティティ間，または，メディア作品間の遷移である 2) 元クエリに周辺語は含まれない[3)]3) 遷移前後で主要語と周辺語の入れ替わりが起きていない

4.2.3 クエリの組に対する CRR の差

クエリの組に対する CRR(Cumulative Reciprocal Rank)[5] の差 ΔCRR はクエリ自動補完の分野においてしばしば用いられる指標である [10, 12][4)]．本論文では ΔCRR を次のように表現する：

$$\Delta CRR = \sum_{d \in D_b} \left(\frac{1}{r(q_b, d)} - \frac{1}{r(q_a, d)} \right) \tag{1}$$

ここで D_b は再検索クエリ q_b によってクリックされうる文書の集合，$r(q_b, d)$ は再検索クエリ q_b に対する文書 d の順位，$r(q_a, d)$ は元クエリ q_a に対する文書 d の順位である．

元クエリ q_a では上位に順位付けできなかった文書を再検索クエリ q_b が上位にできれば ΔCRR は正の値をとる．このとき再検索クエリ q_b はユーザーにとって役に立ったとみなすことができる [12]．

本論文では，$r(q_b, d)$ および $r(q_a, d)$ の順位は所与のクエリに対するクリック先文書の CTR(Click Through Rate) の順位によって算出する[5)]．CTR の算出に用いられた文書の延べ数が各 10 以上，かつ，再検索クエリ側の文書の延べ数が元クエリ側の延べ数の 10%以上，かつ，ΔCRR が 0 より大きく 1.5 以下の場合に元クエリ・再検索クエリの組を正例とする．ΔCRR に対するしきい値の決定方法の詳細は 5.5 節で述べる．

4.3 ラベル未付与事例生成器

次のラベル未付与事例生成器を提案する：

生成器 1 元クエリがあいまいさ回避ページ[6)]と対応するエンティティの名称と一致する事例をラベル未付与とする

生成器 2 クエリの組に対する正例の生成時 (4.2.2 節) に正例と判定されなかった事例のうち周辺語を含まず ID が異なるクエリの組をラベル未付与とする

生成器 3 ΔCRR による正例の生成時 (4.2.3 節) に正例と判定されなかった事例をラベル未付与とする

1） セッションとは，ある特定のユーザーが一定時間内に発行した一連のクエリとそれに伴うユーザ行動のことを指す．

2） 同一 ID を指す再検索クエリが複数ある場合は生起確率の最も高い候補を選択する

3） 元クエリの主要語が空白で区切られている場合，空白の後の文字列は周辺語とはしない

4） 順位に対数を適用した重みづけ [10]，セッション後半のクリック先を利用したスコア補正 [12] などいずれも派生的な ΔCRR の定義をしている

5） 本論文では滞在時間による足切りは行わない

6） https://w.wiki/4dcg

ここで，すでに正例と判定されていた事例についてはラベル未付与とはせず正例とした．

4.4 PU 学習

本論文では次のように PU 学習 [7] を行う．まずラベル未付与の事例に対してラベルを付与する：

1) 正例およびラベル未付与の事例を訓練用とテスト用の 2 つに分割する．2) 訓練用の正例およびラベル未付与の事例を入力とし，ラベル付与確率の回帰器を生成する．3) テスト用の正例に対してこの回帰器を適用し，ラベル付与確率 $g(x)$ の平均 c を求める．4) テスト用のラベル未付与の事例に対して回帰器を適用し $w(x) = p(y = 1|x, s = 0)$ の重みによりラベリングを行う．ここで $w(x)$ は定数 c への依存を持つ．5) テスト用の事例と訓練用の事例を入れ替え，2~4 のステップを行う．

全ての事例にラベルが付与されたら二分割交差検定を行い，各テスト用事例に対して付与された予測確率を順位付けに用いる．

4.4.1 素性

PU 学習の際に用いる素性を表 1 に示す．

表 1: 素性

素性	説明
$Embedding(q_a)$	元クエリ q_a の分散表現 (100 次元)
$Embedding(q_b)$	再検索クエリ q_b の分散表現 (100 次元)
$DiffEmbedding(q_a, q_b)$	q_a と q_b の分散表現の差のベクトル (100 次元)
$CosSimEmbedding(q_a, q_b)$	q_a の分散表現と q_b の分散表現のコサイン類似度
$DeltaCRR(q_a, q_b)$	ΔCRR のスコア
$DeltaCRRFreq(q_a)$	ΔCRR 算出時の q_a のクリック先文書の延べ数
$DeltaCRRFreq(q_b)$	ΔCRR 算出時の q_b のクリック先文書の延べ数
$DeltaCRRRatio(q_a, q_b)$	$DCRRFreq(q_b)/DCRRFreq(q_a)$ の値
$Assist(q_a)$	q_a が発行されたときの検索補助の状況を表現するベクトル．検索窓内の検索補助 (利用，非利用，不明の 3 種類)，検索窓下の検索補助 (利用，非利用，不明の 3 種類)，履歴 (ユーザー履歴経由か否かの 2 種類) の組み合わせ毎の生起回数をベクトル化し，各要素を [0,1] の範囲に正規化したもの (3*3*2 の 18 次元)
$Assist(q_b)$	上記素性の q_b 版 (3*3*2 の 18 次元)
$Assist(q_a, q_b)$	元クエリ q_a および再検索クエリ q_b が連続して発行されたときの検索補助の状況を表現するベクトル (18*18 の最大 324 次元[i])
$DiffAssist(q_a, q_b)$	$Assist(q_a)$ と $Assist(q_b)$ の差のベクトル (18 次元)
$CosSimAssist(q_a, q_b)$	$Assist(q_a)$ と $Assist(q_b)$ のコサイン類似度

[i] 取得期間のログでの生起回数が合計 10 未満である場合は除外

以下，詳細が必要な素性について説明する：

クエリの分散表現　クエリログからトークン数が 2 以上のレコードを抽出し，トークンの生起回数に基づく Shifted Positive PMI の行列 X を生成する．この行列に Randomized SVD[6, 8] を適用し $X = U\Sigma V^*$ を得る．ここで，U,V は直交行列，Σ は特異値の対角行列である．トークンの分散表現に行列 $W^{SVD_\alpha} = U(\Sigma)^\alpha$ を用いる [9][7)8)]．クエリの分散表現はトークンの分散表現の加算により生成する[9)]．

5 評価

5.1 データセット

各データセットの詳細を次に示す：

参照用素性　2021 年 11 月 01 日 ~11 月 30 日までの期間のセッションログを用いて素性を抽出した．

内製のエンティティリンカー　2021 年 12 月 01 日付けのモデル (モデルの訓練には 12 月 01 日付けの内製知識ベース [13] および 12 月 01 日以前の直近 1 年間のヤフー検索のクリックログを利用)

順位付け対象事例　2021 年 12 月 17 日 ~12 月 23 日までの期間に発行された元クエリと再検索クエリの組．4 章で示した操作により順位付け対象事例に対して確率を付与した[10)]．このときの素性は前述の参照用素性を利用した．[11)]

開発・評価事例　順位付け対象事例のうち元クエリに対するエンティティリンカーの推定結果の一位があいまいさ回避エンティティと対応し，かつ，元クエリあたりの再検索クエリの異なり数が 50 以上の事例のみを残した．まず元クエリ 100 事例を非復元抽出し，この元クエリを含む組を評価用とした．評価用として使用されなかった組から ΔCRR が $[0.5, 2.0]$ の範囲の 0.25 刻みで各 10 事例ずつ非復元抽出した．この計 70 事例を開発用事例とした．

5.2 比較手法

比較手法を次に示す[12)]：

RF　Random Forest[4] により PU 学習を行った．Spark MLlib 2.4.6 を使用．Randomized SVD には Criteo Spark-RSVD[13)] を使用．

DCRR　素性抽出用ログの期間に計算した ΔCRR の値を適用した．このときしきい値を 1.5 に設定した．

7) パラメータは $\alpha = 1.0$，negative-sampling の値 $k = 5.0$ に設定
8) クエリ発行時間間隔の制約を一旦解除してトークンの生起回数を計算した．制約下で生起しないトークンの分散表現は除外した．
9) クエリが 1 トークンで構成される場合は，クエリの分散表現はトークンの分散表現と等しくなる．
10) 正例は 1,225,010 事例，ラベル未付与事例は 9,036,884 事例であった．PU 学習前にランダムオーバーサンプリング [1] により両方の数を均等にした．
11) 紙面の都合により本論文では交差検定の結果について取り扱わないが，時系列データなどに対する交差検定の手法はいくつか提案されている．予測対象の直近の区間の素性は利用しないという点で noDepCV[3] に近い手法といえる．
12) 評価事例に対する DCRR 以外の正例生成器 (4.2.1 節，4.2.2 節) の正例生成数はゼロであったため比較手法から除外した
13) https://github.com/criteo/Spark-RSVD

5.3 評価方法

2022 年 1 月 5 日にヤフー検索[14]に対してクエリを発行した．これらの事例に対して次の 3 段階のスコアを付与した：**1.0**： クエリの要件 (2 章) を満たし，かつ，検索結果 1~2 ページ目のいずれかの文書に対応している．**0.5**： クエリの要件を満たさないが，検索結果 1 ページ目の上位 5 件以内の文書と対応している．**0.0**： 上記以外

表 2 にクエリの要件の具体例を示す．

表 2: クエリ要件の具体例．下線部は元クエリ．

クエリの要件	例
適合	“向陽高校 名古屋”，“海鮮三崎港”，“赤い風船 日本旅行”，“花鳥風月 意味”
違反	“ヴァンパイア 歌詞”，“ニャンコ先生 グッズ”，“ウォンツ 店舗”，“スパーク チラシ”

5.4 正例生成器 (ΔCRR) のしきい値の設定

開発事例に対して 3 段階のスコアを付与した (5.3 節)．このとき F 値が最大となったしきい値 1.5 を設定した[15]．

5.5 評価結果

評価事例に対して 3 段階のスコアを付与した (5.3 節)．付与結果をもとにして適合率，再現率，F 値を計測した．表 3 に結果を示す．適合率@3 について DCRR が RF を 12.4 ポイント上回った．再現率@3 および F 値について，RF が DCRR をそれぞれ 12.3 ポイント，4.4 ポイント上回った．

表 3: 適合率，再現率，F 値．Bootstrap 検定 [2] ではいずれの差分も統計的に有意 ($p < 0.05$) であった．

手法	適合率@3	再現率@3	F 値
DCRR	**0.581 (+0.124)**	0.298	0.394
RF	0.457	**0.421 (+0.123)**	**0.438 (+0.044)**

表 4 に順位付け結果の例を示す．RF はクエリ “花鳥風月” に対して知識外の花鳥風月 (ビール) も含めて順位付けできた．また，クエリ “山頭火” に対して周辺語を含まない部分一致の再検索クエリである “種田山頭火” も含めて順位付けできた．

表 4: 順位付け結果．下線かつ太字はスコア 1.0，下線のみはスコア 0.5，下線無しはスコア 0．左から順に 1，2，3 位

元クエリ	順位付け結果
花鳥風月	**DCRR**： “**花鳥風月 意味**” **RF**： “**花鳥風月 意味**”，“花鳥風月 歌詞”，“**花鳥風月 ビール**”
山頭火	**DCRR**： “**山頭火 ラーメン**” **RF**: “**山頭火 ラーメン**”, “**山頭火 カップラーメン**”, “**種田山頭火**”

14） https://search.yahoo.co.jp
15） このときの F 値は 0.735

5.6 素性の重要度の分析

Random Forest[4] の出力した分散に基づく素性の重要度の上位 25 件を図 2 に示す．上位を独占したのは ΔCRR 関連の素性であった．ΔCRR 自身の値 $DeltaCRR(q_a, q_b)$ よりもその算出に関わったクリック先の文書数の比率 $DeltaCRRRatio(q_a, q_b)$ の方が重要という結果となった．この素性は暗黙的に候補間での相対的な再検索クエリの人気度を包含しているため，これが重要だったのではないかということが考えられる．次点は検索補助関連の素性という結果となった．上位の $Assist(q_a)_{16}$,$Assist(q_b)_{14}$ は同一素性でもベクトルの位置が異なっている．元クエリと再検索クエリでは注目するべき検索行動が異なるということが示唆される．クエリの分散表現関連の素性は上位 25 件内に確認できなかった．この理由としては，正例生成器において ΔCRR とエンティティリンカーの判定の論理和を利用したため，本来クエリの意味の観点からは正例に算入するべきではない事例 (e.g., “XXX 店舗” の構成の再検索クエリを含む組) が一律に正例として混じってしまい，回帰器の学習時にこの素性が利用される場面が限定的になってしまったのではないかと考えられる．

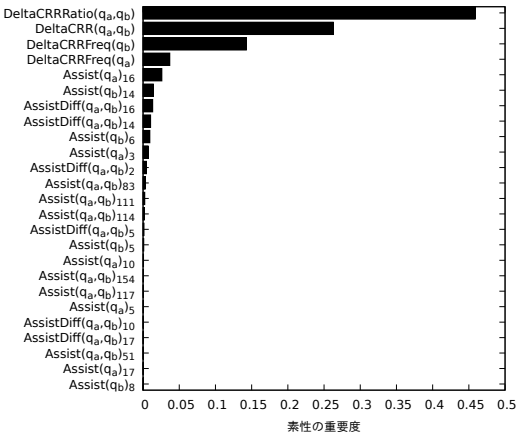

図 2: 素性の重要度．添え字はベクトルでの位置 (0-base)．

6 おわりに

本論文では複数の正例生成器とラベル未付与事例生成器を組み合わせることで訓練事例を生成することを提案した．ΔCRR を単独で用いる場合よりも，複数の素性を用いた Random Forest[4] による順位付けのほうが F 値が 4.4 ポイント向上することを示した．

参考文献

[1]Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, Vol. 6, No. 1, pp. 20–29, 2004.

[2]Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 995–1005. Association for Computational Linguistics, 2012.

[3]Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, Vol. 191, pp. 192–213, 2012.

[4]Leo Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[5]Allison JB Chaney, David M Blei, and Tina Eliassi-Rad. A probabilistic model for using social networks in personalized item recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 43–50, 2015.

[6]Paul G Constantine and David F Gleich. Tall and skinny qr factorizations in mapreduce architectures. In *Proceedings of the second international workshop on MapReduce and its applications*, pp. 43–50, 2011.

[7]Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

[8]Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, Vol. 53, No. 2, pp. 217–288, 2011.

[9]Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, Vol. 27, pp. 2177–2185, 2014.

[10]Umut Ozertem, Olivier Chapelle, Pinar Donmez, and Emre Velipasaoglu. Learning to suggest: a machine learning framework for ranking query suggestions. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 25–34, 2012.

[11]Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, pp. 771–780. ACM, 2010.

[12]Milad Shokouhi, Marc Sloan, Paul N Bennett, Kevyn Collins-Thompson, and Siranush Sarkizova. Query suggestion and data fusion in contextual disambiguation. In *proceedings of the 24th international conference on world wide web*, pp. 971–980, 2015.

[13]Tomoya Yamazaki, Kentaro Nishi, Takuya Makabe, Mei Sasaki, Chihiro Nishimoto, Hiroki Iwasawa, Masaki Noguchi, and Yukihiro Tagami. A scalable and plug-in based system to construct a production-level knowledge base. In *Proceedings of the 1st International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs co-located with the 25th ACM SIGKDD International Conference on Knowledge Discovery &Data Mining*, 2019.

[14]豊田樹生, 小松広弥, 熊谷賢, 菅原晃平. ウェブ検索クエリのための部分一致文字列に対するエンティティ名称予測モデルの提案. 言語処理学会第 27 回年次大会発表論文集, pp. 590–594, 2021.

[15]豊田樹生, 土沢誉太, 築地毅, 菅原晃平, 野口正樹. dishpam: A distributable seeded hierarchical pachinko allocation model. 言語処理学会第 26 回年次大会発表論文集, pp. 217–220, 2020.

[16]豊田樹生, 夜久真也, 石川葉子, 土沢誉太, Kulkarni Kaustubh, Bhattacharjee Anupam, 宰川潤二. ウェブ検索クエリに対する周辺語を考慮した教師なしエンティティリンキング. 言語処理学会第 25 回年次大会発表論文集, pp. 81–84, 2019.