2B3-1

ウェブ情報を用いたエンティティのランキング学習に関する研究

Learning to rank entities using Web information

森 純一郎*1　松尾 豊*1
Junichiro Mori　Yutaka Matsuo

*1東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

In this research, we aim to develop a method for learning to rank entities using Web information. Most of existing methods rely on solely IR-based features to rank entities. However, other features may be useful for ranking entities. In particular, a lot of additional features such as document-based, structure-based, and graph-based can be acquired from heterogeneous Web sources including search result, Web page, Blog/SNS, other databases. To find the best combination of the features in an automatic manner, we employ a supervised learning approach to rank entities and relations using Web information. Our method contributes to entity search by providing the general ranking framework of entities and relations on the Web.

1. はじめに

ウェブ上の膨大な情報を検索可能にする検索エンジンおよび実世界情報を投影したブログ、マイクロブログ、ソーシャルネットワークサービスといったソーシャルメディアの普及により、近年のウェブ情報検索は従来のドキュメントを対象とした検索から人物、組織、場所と行った実世界のエンティティを対象とした検索へと変化している。実際に、検索エンジンのクエリーの多くは人物名であり、最近では人物や企業といった特定のエンティティの検索に特化したサービスが運用されるようになってきている。

エンティティ検索における課題の一つは、クエリーに応じてエンティティをどのようにランキングするかである。従来の研究においては、エンティティのランキングは特にエキスパート発見の分野で研究がなされている。エキスパート発見における、エンティティ（人物や組織）の従来のランキング手法は対象となるエンティティをテキストの特徴量によって重み付けしランキングするものである。これらの手法の基本的な考え方は、クエリーと関連のあるより多くの語・テキストとエンティティが関係していれば、クエリーに対してエンティティは深く関連しているというものである。その際には、エンティティと語・テキストとの関係が保証されなければならない。例えば従来のエキスパート発見タスクが対象とする企業内文書 (プロフィール、メール等) であれば、エンティティとテキストの関連は明確であるが、ウェブ上の情報、例えば検索結果、ウェブページ、ブログ、ソーシャルネットワークサイト、の場合はエンティティとテキストと関連が曖昧である。そのため、ウェブ上の情報に基づいてエンティティのランキングを行なう場合は、エンティティと情報を関連づけるランキングに有効な属性を特定することが必要となる。

本研究では、ウェブを対象としたエンティティ検索を目的とし、教師あり学習に基づくウェブ情報を利用したエンティティランキング手法を提案する。提案手法においては、検索クエリーに対して様々なウェブ上の情報を収集し、それらの情報から得られた属性を元にエンティティのランキング学習を行う。本論文では特に、人物エンティティを対象に、実際のエンティティ検索システムから得られたデータを利用して、ウェブ情報を利用したエンティティのランキング手法について例を示す。

連絡先: 森純一郎，東京大学大学院工学系研究科，東京都文京区弥生 2-11-16 工学部 9 号館 119 号室，03-5841-1161，jmori@ipr-ctr.t.u-tokyo.ac.jp

2. 関連研究

検索クエリーに応じてエンティティをランキングする手法は、主にエキスパート発見の分野で研究がなされてきた。特に、近年では情報検索の視点からエキスパート発見に取り組む研究が行われてきており、TREC や INEX などの情報検索のワークショップにおいては、エキスパート発見あるいはエンティティのランキングがタスクとして取り入れられている。これらの手法は、従来のテキスト検索をエンティティ検索に適用したアプローチに基づくもの [Balog 09] であり、基本的には統計的言語アプローチに立脚したテキストの特徴量 (TFIDF, BM25，共起情報、言語モデルなど) によりエンティティのランキングを行なっている。一方で、検索クエリーとエンティティに関わる複数の特徴量を属性として用いて学習によりエンティティのランキングをする研究も行なわれている [Serdyukov 08, Hu 06]。しかしながら、これらの学習で扱う属性は局所的な情報であり、ウェブ上の大規模かつ異種混在の情報からエンティティのランキングを行なうためにどのような属性を生成し組み合わせるかは課題である。

3. ウェブ情報を用いたエンティティのランキング学習

3.1 人物検索エンジン

ウェブ上におけるテキスト検索からエンティティ検索へという流れの中で、最近では人物や企業といった特定のエンティティの検索に特化したサービスが運用されるようになってきている。特に、人物検索においては多くのサービスが公開されており、日本国内においては *SPYSEE* *1 という人物検索サイトが多くのアクセスを集めている。*SPYSEE* は、数十万人にわたる人物情報をウェブから自動抽出し、人物ごとのページを自動生成している (図 1)。ユーザは任意のクエリーで人物検索することで各人物のページにたどり着き、その人物情報を得ることができる。

*1　http://spysee.jp

図 1: 人物検索サイト SPYSEE

3.2 人物検索エンジンにおけるランキングの学習

本研究においては、*SPYSEE* におけるユーザの検索クエリーと実際にその検索クエリーからユーザがたどり着いた検索ページを人物検索におけるランキングの正解データとして任意のクエリーに対する人物ランキングの学習を行なう。ユーザの検索クエリー q から人物エンティティ e のページにたどりついた回数を $land(q,e)$ とする。この時、ある検索クエリー q に対して人物エンティティ e_i と e_j の間に $land(q,e_i) > land(q,e_j)$ なる関係が存在する時に、q に関して e_i は e_j よりもランクが高いとする。

本研究のランキング学習においては、属性として検索エンジンのヒット件数を利用した次の特徴量を用いる。

- 属性 1：クエリ q の単独ヒット件数 ($|Q|$)
- 属性 2：エンティティ e の単独ヒット件数 ($|E|$)
- 属性 3：クエリとエンティティの共起ヒット件数 ($|Q \cap E|$)
- 属性 4:クエリとエンティティのシンプソン係数 ($|Q \cap E|/min(|Q|,|E|)$)
- 属性 5：クエリとエンティティのダイス係数 ($2 * |Q \cap E|/(|Q| + |E|)$)
- 属性 6：クエリとエンティティのジャカード係数 ($|Q \cap E|/|Q \cup E|$)

ここで、$|A|$、$|A \cap B|$、$|A \cup B|$ は、それぞれ"A"、"A" "B"、"A" OR "B"、で検索エンジンを検索した時のヒット件数を表す。これらのヒット件数に基づく特徴量は、TFIDF などエンティティに関する従来の局所的なテキストの特徴量に対して、大規模なウェブの情報を用いた大域的な特徴量 [Matsuo 02] である。これらの特徴量を属性として、クエリーログから得られた正解データをもとに任意のクエリーに対する人物ランキングの学習を行なう。学習には RankingSVM [Joachims 02] を用いる。

4. 実験

人物検索サイトではさまざまなクエリーが検索されている。エンティティをどのようにランキングするかは、検索クエリーの種類と深く関連している。例えば、職業を表すような検索クエリーの場合は、ユーザの要求は特定の分野におけるエキスパート発見である場合が多い、また組織や団体を表すような検索クエリーの場合は、ユーザは著名人を探している場合が多い。本実験においては、特にクエリー中に" 大学" を含むクエリーを対象として *SPYSEE* から正解データを取得したクエリーの種類は 118 であり、1654 の学習データを元に人物ランキングの学習を行なった。RankingSVM は線形カーネルを用い、正則化には L1 を用いた。

表 1: 人物のランキング学習における属性重み

属性 1	0.0
属性 2	2.3310041
属性 3	0.77327365
属性 4	-1.7141689
属性 5	0.54906631
属性 6	0.84371138

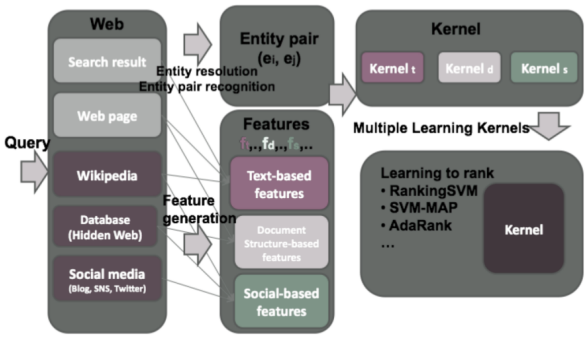

図 2: ウェブ情報を用いたエンティティのランキング学習手法

表 1 は、学習モデルにおけるそれぞれの属性重みを表している。クエリーによらず、人物のウェブ単独ヒット件数 (属性 2) は、その人物のランキング学習に寄与していることがわかる。さらにクエリーと人物の共起ヒット件数 (属性 3) やジャカード係数 (属性 6) も、ランキング学習へ寄与している。一方で、クエリーと人物のシンプソン係数はランキング学習において負の寄与を示している。

5. おわりに

本研究では、ウェブを対象としたエンティティ検索を目的として、教師あり学習によるエンティティランキング手法を提案し、実際の人物検索エンジンを元に基本的な実験を行なった。本実験で用いたウェブヒット件数に基づく属性は、テキスト・言語に基づく属性である。一方で、膨大で多種多様なウェブ情報から得られるクエリーとエンティティの属性はテキスト情報にとどまらず構造情報、リンク情報、さらにはソーシャルメディア等から得られるソーシャルな情報など多岐に渡る。今後は図 2 に示すように、これらの情報から得られるさまざまな特徴量に基づく属性を設計し、それらを統合したカーネルを構築することでエンティティのランキングを学習する手法の開発を行なう。これにより、ウェブ上の情報からエンティティをランキングするのに有効な属性の生成と評価を行なう予定である。

参考文献

[Balog 09] K. Balog, L. Azzopardi, and M. Rijke: A language modeling framework for expert finding, Information Processing and Management, Vol. 45, 2009.

[Hu 06] G. Hu, J. Liu, H. Li, Y. Cao, J.Y. N, and J. Gao: A Supervised Learning Approach to Entity Search, AIRS, 2006.

[Serdyukov 08] P. Serdyukov, R. Aly, D. Hiermstra: Using the Global Web as an Expertise Evidence Source, TREC Enterprise Track, 2008.

[Joachims 02] T. Joachims: Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.

[Matsuo 02] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, Mitsuru: POLYPHONET: an advanced social network extraction system from the web, Proceedings of the 15th international conference on World Wide Web, 2006,