論文の内容の要旨

Latent Relational Web Search Engine Based on

the Relational Similarity between Entity Pairs

（エンティティペア間の関係類似度を利用するウェブ潜在関係検索エンジン）

氏名　　グェン トアン ドゥク

The World Wide Web contains a huge number of Web pages which refer to numerous semantic relations. When a user wants to search for an entity in a specific semantic relation using a keyword-based Web search engine, the user must formulate a query with some keywords related to the entity and the relation. The user then inputs this query into the keyword-based Web search engine to retrieve a set of text snippets which the user must read to find out the answer. Moreover, when one does not explicitly know appropriate keywords to formulate a query, one can not get answers by using keyword-based Web search engines. With the growing number of entities and semantic relations on the Web, Web search engine users frequently face with such situations. Therefore, new entity retrieval paradigm based on semantic relations between entities is required to alleviate this problem. In this thesis, we study the problem of latent relational search, a novel entity retrieval method that enables Web search engine users to directly retrieve appropriate entities in an implicitly stated semantic relation. Specifically, given a latent relational search query {(A, B), (C, ?)}, in which A, B, C are entities, a latent relational search engine is expected to retrieve a list of entities L containing candidate answers to fill in the question mark (?) in the query. In the list L, each entity D satisfies the condition that the semantic relation between A and B is highly similar to that between C and D. For example, given the query {(Japan, Tokyo), (France, ?)}, a latent relational search engine is expected to retrieve and rank the entity ``Paris'' as the first answer in the result list, because the relation between Japan and Tokyo is highly similar to that between France and Paris.

To perform latent relational search on the Web, one must overcome several challenges: discovering entity pairs to build an index for high speed retrieval, exploring and representing the semantic relations between entities, and ranking the candidate answers according to the degree of relational similarity between the candidate entity pairs and the input pair. We propose a method for extracting entity pairs from a text corpus to build an index for a high speed latent relational search engine. Following previous work on relational similarity measuring algorithms, we represent the relation between two entities in an entity pair using lexical patterns of the context surrounding the two entities. We propose a lexical pattern extraction algorithm which enables the search engine to precisely measure the relational similarity between two entity pairs and therefore to accurately rank the result list of a latent relational search query. Different from previous work on latent relational search, the proposed retrieval model allows supporting sentences to be retrieved as evidences for each result. These evidence sentences provide the users of the

search engine with further knowledge concerning the common semantic relations between the input entity pair and each retrieved candidate entity pair.

Moreover, we propose cross-language latent relational search, an advanced latent relational search paradigm that allows answering the query {(A, B), (C, ?)} when the input pair (A, B) is written in another language from the language of the entity C. To capture the similarity between relations across languages, we must transfer the meaning of lexical patterns from one language to another. We propose a novel lexical pattern clustering algorithm to recognize paraphrased lexical patterns across languages, thereby effectively ranking candidates and retrieving evidence sentences for cross-lingual queries.

We evaluate the proposed search engine on both monolingual query sets and English-Japanese cross-lingual query sets. The experimental results show that, the proposed method outperforms existing latent relational search engines on monolingual query sets. The search engine also achieves a moderate Mean Reciprocal Rank (MRR) on cross-lingual latent relational search query sets. Importantly, for the majority of cross-lingual queries, the search engine retrieves supporting sentences that are semantically similar in two different languages. This implies that the results of the search engine can be used for building parallel corpora or for supporting human translators. In particular, when evaluating with an ideal corpus, the proposed search engine retrieves the correct answer in the Top 1 ranked result for 95% of monolingual queries in English and 88% in Japanese. When evaluating with Japanese - English cross-language latent relational search queries, the proposed method achieves an MRR of 0.605 while requiring an average query processing time less than 10 seconds, which is acceptable for normal search engine sessions. Finally, we show that the proposed model can be applied to build a large-scale latent relational search engine with real-world corpora. Specifically, we use seven million articles in the English and Japanese Wikipedia data dumps to build an index for the search engine and use the search engine to answer several sophisticated questions in the INEX 2008 Entity Ranking task. The results show that, the search engine was able to answer 15 (out of 35) queries in monolingual mode, where as, in cross-lingual settings, the number of successfully answered questions was 12 (out of 35). This demonstrates that the proposed system could be used for answering sophisticated questions concerning entities and relations on the Web.

From these results, this work reveals the possibility of latent relational search as a next generation information retrieval and question answering paradigm on the Web.

日本語の概要:

膨大な WWW 情報空間の中には，様々なエンティティとそれらの関係が多数，潜在的に記述されている．我々は従来のキーワードベースの Web 検索エンジンを利用することでキーワードを含む文書は検索できるが，エンティティ間の関係を検索することはできない．このため，エンティティ間の関係に基づいた検索を実現するため，潜在関係検索という新しい検索パラダイムが検討されてきた．潜在関係検索とは，与えられたエンティティペア(A, B) とエンティティ C に対して，(A, B) と(C, D) が類似関係を持つようなエンティティ D を探す検索のことである．

例えば，潜在関係検索エンジンに {(Japan, Tokyo), (France, ?)} というクエリが入力されると，``Paris"を最初にランキングした結果リストを返す．これは，Japan と Tokyo との関係が France と Paris との関係と類似するからである．WWW 空間の情報爆発が顕著となった昨今においては，検索エンジンにおいて適切なキーワードと検索クエリを考え出すことが一般ユーザにとって困難となりつつあり，単純なキーワードを含む Web ページの検索だけでは現状に対応するのには限界がある．そこで本研究では潜在関係検索に着目し，高速かつ高精度な検索を，多言語な Web 空間上で行う手法を提案する．

本研究ではまず，高速な潜在関係検索を行うために，エンティティペアを Web から発見・抽出する手法と，抽出されたエンティティペアに対する Index の構築手法を提案する．また，高精度な潜在関係検索エンジンを実現するため，エンティティ間の関係を周辺文脈の語彙パターンで表現し，精度の高い関係類似度計算アルゴリズムを利用し，Index を用いて検索結果のランキングを行う．提案手法は従来の潜在関係検索の手法とは異なり，検索の過程で発見された入力エンティティペアと候補ペアとの共通の語彙パターンを使用して，候補エンティティだけを結果として出力するのではなく，そのエンティティがなぜ出力されるかという根拠の文も出力する．

更に本研究では，多言語のテキストを利用する言語横断型の潜在関係検索を提案する．言語横断型の検索では，入力ペア (A, B) と入力されたエンティティ C が異なる言語で書かれている場合も検索可能である．これにより，候補エンティティ D が記述された文と，入力ペア(A, B) が記述された文が異なる言語の場合も検索ができ，検索に用いられる根拠の文の数と検索可能なエンティティの範囲を広げるができる．言語横断型の潜在関係検索を実現するためには，意味関係を特徴付ける語彙パターンを，入力ペアの言語から出力ペアの言語へとマッピングする必要がある．本研究ではこの意味マッピングの問題を解決するために，類似する語彙パターンを言語横断的に認識する手法を提案する．具体的には，新しい語彙パターンクラスタリング アルゴリズムを提案し，言語横断的に語彙パターンをクラスタリングする．これにより，言語横断型のクエリを処理でき，意味の近い根拠文を異なる言語から取り出すことができる．

評価実験では，まず小規模なテキストコーパスを使い，各関係タイプにおける提案手法の性能を評価する．具体的には，8 種類の関係で平均逆順位 (MRR) と，正解を Top 1 の結果にランキングできるクエリの割合を調べることで評価を行う．その結果，95%の英語の単一言語検索クエリにおいて，正解を Top 1 にランキングすることができた．また日本語のクエリでは，88%のクエリで正解を Top 1 にランキングすることができた．これは提案手法が，従来の単一言語の潜在関係検索の手法よりも精度の高い検索ができることを示している．一方，言語横断型のクエリセットでは，平均逆順位 (MRR) で 0.605 の値を達成した．この実験では根拠文として，言語は異なるが意味は類似した文を取得できた．また，平均クエリ処理時間は 10 秒以内で

あり，実用レベルのクエリクエリ処理時間であった．これにより，提案システムにはユーザの翻訳作業を支援できる可能性があることが示された．

次に，大規模なコーパスにおいて実用レベルの潜在関係検索エンジンを実装し，INEX 2008 Entity Raking タスクを用いてその性能を評価した．提案した潜在関係検索エンジンは，単一言語の検索において，INEX の 35 の質問の中で 15 の質問に対して正解を出力できた．また，言語横断のクエリが入力された場合では，35 の質問のうち，12 の質問に対して正解を出力できた．これにより，提案システムが実用レベルで質問応答システムとして使えることが示された．

上記の成果を通して本研究では，潜在関係検索が次世代の情報検索パラダイムになり得ることを明らかにする．