博士（工 学） スコブロン マーティン グジェゴシュ

学 位 論 文 題 名

A Web Based Approach to Factoid and Commonsense Knowledge Retrieval

（WWWを利用した事実及び常識に関する知識の検索）

学位論文内容の要旨

In recent years, the amount of machine readable resources has been rapidly increasing, including the textual content of the WWW. While the amount of information increases, it has also become more difficult for an average user to access specific and reliable information. Consequently, the usability of this resource is limited, especially for users who do not possess sufficient knowledge and/or experience with using the currently available solutions. Since language along with its communication function is used to represent knowledge and human beings' recognition of the world, one can perceive the growing size of textual resources as an asset usable for providing commonsense knowledge to computer systems. Without such knowledge machines cannot reason in a way similar to human beings, which slows down the spread of AI systems. The dissertation addresses these two issues by describing and evaluating methods for factoid and commonsense knowledge retrieval from the WWW.

Chapter 1 presents Internet as a source of information and knowledge usable both for human beings and computer systems. Then the research motivation is presented with an overview of proposed methods and solutions, along with their applications to the Question-Answering, Interactive Clustering-Driven Question-Answering and Automatic Knowledge Retrieval Systems.

Chapter 2 describes methods applicable to question-answering systems including question classification, query formation and answer candidate extraction and verification. After the introduction that provides background on question-answering, the chapter presents in detail the question classification task and the applicably of question category information to various parts of the Question-Answering System. The proposed question classification method uses a Support Vector Machines based classifier trained with an extended set of features applied to a fine-grained taxonomy that includes 50 question category types. The described method provided the highest classification accuracy rate reported in literature, outperforming other approaches that used the same training and test data.

The basic idea that motivates the Query Generation Patterns method is that for questions that belong to the same question category and have a similar syntax, the same generation rules can be used to form a set of queries that retrieve an answer-rich set of documents. The Query Generation Patterns allows the Question-Answering System to automatically acquire knowledge on how to form reliable queries. It provided information on an optimal combinations and modifications of question words; possible extensions of a query with non-question words, that can be applied to questions from a given questions category and syntax.

The answer candidate extraction and verification processes extensively use question category information to provide constraints on the scope of plausible answer candidates. This section presents the creation and application of the fine-grained Entity Recognition tool, as well as various question category dependent answer candidates extraction methods and search strategies; including the question category reliant selection of highly applicable and reliable online sources.

The motivations to present the Interactive Clustering-Driven Approach to Question Answering includes the observation that a natural communication between human-beings is not limited to a single question-answer pair. It frequently involves interlocutors interacting to further specify sought-after information, reconfirming or rephrasing an original question. Additionally, a human being possesses intuition, commonsense or partial knowledge on a topic that can be used to support the Question-Answering System. A meaningful system-user interaction, based on initially retrieved set of documents provides the means to adjust the search space and to select an answer rich set of documents closely related to a question. It also addressed users expectations concerning naturalness and intuitiveness of accessing sought-after information.

Chapter 3 presents the method for automatic knowledge retrieval from the Internet resources. After the motivation for this research is presented, the chapter describes other projects that aim at the creation of a knowledge base of assertions or concepts about everyday world. The goal of the presented method is to automatically create entries to a knowledge database, expressed in the natural language form, similar to the ones provided by the volunteer contributors in a projects like Open-Mind Commonsense (OMCS). Since only a small fraction of the statements accessible on the Web can be treated as valid knowledge concepts methods for their filtering and verification were considered, based on similarity measurements with the concepts found in the manually created knowledge database.

Chapter 4 describes the application of the presented methods to Question-Answering and Knowledge Retrieval Systems. The evaluation of the methods proposed for the question-answering was performed using questions from the QA Track of the Text Retrieval Conferences (TREC). The experiments demonstrate the effectiveness of the proposed methods and their influence on the usability, behavior and overall accuracy of the Question-Answering and Interactive Question-Answering Systems. The application of the Knowledge Retrieval method proved that the system was capable of automatically discovering and filtering knowledge concepts in a user-selected domain with high accuracy. Some of the automatically retrieved knowledge concepts provided semantic equivalents of the statements that were manually input to the OMCS. Other statements, while including more details compared to the OMCS entries, could also become a part of a knowledge database. The results also confirmed that the system was able to retrieve high quality knowledge concepts, even for the terms that were not described in the knowledge database built by mass collaboration of Internet users.

Chapter 5 presents overall conclusions and discussion on the future development of solutions for accessing sought-after information and knowledge from machine readable resources, as well as applicability of commonsense knowledge to computer systems for smoothing the human-machine interaction.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　北　島　秀　夫
副　査　教　授　山　本　　強

学位論文題名

A Web Based Approach to Factoid and Commonsense Knowledge Retrieval

(WWWを利用した事実及び常識に関する知識の検索)

近年，WWW 資源に存在する機械可読資源の量が急速に増大している．著者はこれまで検索エンジンの使用がユーザの直感に適合したものではなく，経験の浅いユーザが信頼性の高い適切な検索結果を得ることができないため，必要なデータにたどり着くまでに膨大な時間を要するという問題を解決するために研究を行ってきた．著者は本論文で自然言語がコミュニケーション機能のみではなく，世界知識及び世界認識を表現するために自然言語が使用されているので，WWW やコーパスなどのテキスト・リソースの増大が計算機システム用の常識的知識の知識源として利用できることを主張している．

著者は事実および常識的知識の WWW からの抽出の研究を行い，上記で述べた問題を解決し得る手法を発表し，その手法を用いた実験システムを開発し，性能評価実験を行った．著者が提案した手法は，インタラクティブなクラスタリング手法（Interactive Clustering Driven Approach），クエリ生成パターンを用いたクエリ生成手法（Query Formation based on the Query Generation Patterns Method），素性拡張セットを用いた SVM による質問分類手法（Support Vector Machines Based Question Classification Method with the Extended Set of Features）であり，これらの手法をテキスト検索，応答候補抽出と選択の処理に応用した．また，著者は WWW 上の常識的知識及び一般信念を計算機システムが利用できる形に変換する手法を開発した．提案したこれらの手法は Web に基づく質問応答システム(Web Based Question Answering System)，インタラクティブなクラスタリング手法による質問応答システム（Interactive Clustering Driven Question Answering System）及び

Webに基づく知識獲得システム（Web Based Knowledge Retrieval System）に応用され，実験により性能評価を行った．

著者が提案した質問分類手法は従来手法に比べて高い精度を得ている．また，質問応答システムにおいてはクエリ生成パターンを使用することによって自動的な情報抽出が可能になり，それが正しい応答を得るための適切なクエリであることが示された．インタラクティブなクラスタリング手法による質問応答システムを使用することによりユーザの直感，常識などを質問応答システムで利用できることを示した．質問応答システムの利用時の最初の段階で得られたドキュメントを用いたシステムとユーザのインタラクションにより探索空間の縮小を行い、質問と密接に関連する応答を含む文書を獲得できることを著者が実証した．上記で述べた手法は国際的な情報検索のコンテストで使用された質問応答システムのためのテストデータによってその性能評価実験が行われた．この実験結果により，著者によって提案された手法の有効性及び Web に基づく質問応答システムとインタラクティブなクラスタリング手法による質問応答システムの使いやすさ，精度の高さを実証した．

また，著者はインターネット・リソースから知識を自動的に抽出する手法を提案した．提案した手法の目的は人手で作成され自然言語で表現されている常識データベースである Open-Mind Commonsense (OMCS)の自動生成である．提案手法に基づくシステムは WWW から高精度な知識概念を抽出し，フィルタリングを行う．多くの場合には自動的に抽出された知識概念はOMCS の人手により作成されたエントリーと同じ意味を表している．さらに提案手法に基づくシステムによって得られた知識概念は OMCS より詳細であり，かつ OMCS に存在しないものも抽出できることが示された．

以上を要約すると，著者はインタラクティブなクラスタリング手法，クエリ生成パターンを用いたクエリ生成手法，素性拡張セットを用いた SVM による質問分類手法を提案し，Web に基づく質問応答システム，インタラクティブなクラスタリング手法による質問応答システム及び Web に基づく知識獲得システムに応用し，高い精度を得た．本研究による情報メディア工学，自然言語処理工学への貢献は大なるものがある．よって，著者は博士（工学）の学位を授与される資格あるものと認める．