1X-06 質問応答システムのための新しい文書検索手法の提案

松村真宏 †*　大澤幸生 ‡　石塚満 †

† 東京大学大学院工学系研究科
‡ 筑波大学経営システム科学専攻

Abstract

ユーザの入力する質問に回答する質問応答システムにおいては、予め全ての質問に対する回答を用意しておくことは難しい。本稿では、一文書でユーザの質問が解消されない場合には複数の文書を組み合わせて回答を作成する新しい文書検索手法について述べる。

1 はじめに

ユーザの入力する質問に回答する質問応答システムにおいては、予め全ての質問に対する回答を用意しておくことは難しい。そこで本稿では、一文書でユーザの質問が解消されない場合には、複数の文書を組み合わせてユーザの質問を満たすような回答を作成する新しい文書検索手法について述べる。

2 関連研究

ユーザの質問から回答を導く FAQ Finder System[1] は代表的な質問応答システムである。また、WWW 上の文書（Web ページ）を検索するサーチエンジンも、検索キーワードがユーザの質問を表す場合には質問応答システムと見なすことができる。しかし、これらのシステムでは複数の文書を組み合わせた効果は考慮していないため、適切な回答が一文書として存在しない場合の検索精度は悪い。

ユーザの質問を満たす必要最小限の文書集合を出力するシステムには NaviPlan[3] と AAS[2] がある。NaviPlan は、ユーザが理解したい概念（目標概念）を受け取り、その概念理解に有用な文書集合（プラン）を返すシステムであり、目標概念を基礎から理解させることを狙ったシステムである。NaviPlan では直列に連なった文書集合が回答となるため、順番通り読み進めることが前提となる。しかし、NaviPlan は目標概念に直結する文書がなければプランが得られないため、目標概念の 6 割ほどしかプランが生成されない。一方、AAS はユーザの質問を部分的に満たす文書を組み合わせることで、回答が一文書として存在しない場合にもユーザの質問を十分に満たす回答を作り出すシステムである。AAS では出力する文書集合は並列関係となるので、文書を読み進める順番は問わない。しかし、AAS は得られた文書集合を理解するのに文書以外の知識を要するため、回答がユーザの知識では理解できない場合もある。

3 提案手法

NaviPlan は狭く深い理解、AAS は広く浅い理解を導く。しかし、ユーザの持っている知識は十人十色であり、またユーザの理解したがっている知識にも差があるため、NaviPlan と AAS の両方の特徴を兼ね備えることが理想である。そこで本稿では、AAS の処理を拡張することにより、ユーザを広く浅い理解のみならず、深い理解まで導くことを狙う。

AAS では、ユーザの質問をゴールとし、データベースから作成した背景知識、仮説知識に基づいてコストに基づく仮説推論によりコスト[1]最小となる文書集合 1 組だけを選んで回答としている。紙面のスペースの関係で AAS のアルゴリズムの詳細は [2] に譲るが、AAS は文脈を共有するよう文書集合を選ぶので、互いに説明が不足している部分を補い合う効果が得られ、その結果ユーザにとって読みやすく有益な文書集合が回答として得られる。

しかし、その回答文書が理解できない場合には、その文書を説明するような文書を新たに加えてユーザを理解に導くことが重要となる。AAS の出力する文書集合は、ユーザが同時に読むべき文書の関係を表しており、この関係はユーザの興味に応じて動的に生成されるものである。そこで、コストが比較的少ない解を複数集めてこの関係を眺めることにより、ユーザの興味をより網羅的に反映した文書ネットワークを考えることができる。本稿では、このようにして得られる文書ネットワークからユーザの興

*A Proposal of Combination Retrieval for Question Answering System, Naohiro Matsumura, Faculty of Engineering, the University of Tokyo.

[1] AAS ではユーザが知っておくべき外部知識の数をコストとしている。

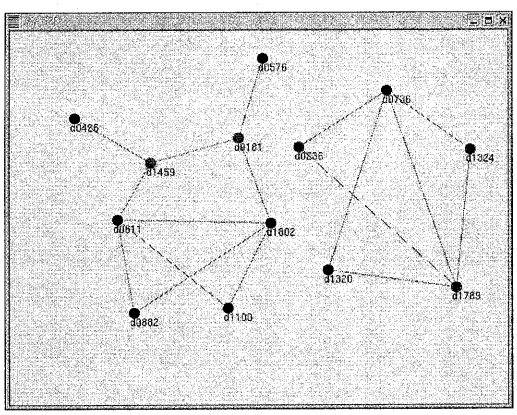

図 1: 検索語が"alcohol, fat, calorie"の時に得られる文書ネットワーク。

味に応じた最適な文書の組み合わせを回答として取り出すことを考える。

本稿で提案するアルゴリズムは次のようになる。

Phase1.AAS フェーズ 1 AAS と同様の処理によりゴール、背景知識、仮説知識を作成し、コスト最小となる回答（文書集合）を得る。

Phase2.AAS フェーズ 2 **Phase1** により得られた回答がくり返し得られないように inc 条件を背景知識に追加し、所定の数の回答（本稿では 10 個）が得られるまで **Phase1** をくり返す。但し、ネットワークを適度に成長させるため、同一の文書が 4 回以上登場する場合はその文書を inc 条件に加える。

Phase3. 文書ネットワークの構築 各回答がそれぞれ 1 つのクラスタになるように文書間にリンクを張ることにより、ユーザの興味を網羅的に反映した文書ネットワークが得られる。

Phase4. 回答の探索 文書ネットワークを辿ることにより必要な文書を選択し読み進める。

4 実験

提案手法を逐次的に実行できる環境を Sun Enterprise450 上に構築し、実験による評価を行った。実験に用いたデータベースは、コロンビア大学のヘルスカウンセリングセンターにあるオンラインデータ[2]約１８００件である。

お酒に含まれる脂肪やカロリーが体に与える影響を知りたくて、検索語として"alcohol, fat, calorie"を与えた時の文書ネットワークを図 1 に示す。d1459 と d0181 で示されるノードが従来の AAS で得られる回答である。d1459 は「カロリーが不足するとタンパク質を燃焼させてエネルギーを得るためタンパク質が不足するが、タンパク質が不足すると疲労の回復が遅れたり病気にかかりやすくなるなどの弊害が起こる。」に関する文書、d0181 は「過剰なアルコールの摂取は肝臓や心臓に深刻なダメージを与える。」に関する文書である。これらの文書を読むことによりユーザの興味は「カロリー不足がもたらす弊害」もしくは「適度なアルコールの摂取量」に移ると考えられる。しかし、文書ネットワーク上で d0181 に繋がっている d0576 は理想的な一日のカロリー摂取量について書かれた文書、また d1459 に繋がっている d0611 はアルコールの適切な摂取量について書かれた文書であるため、ユーザの興味はこれらの文書により十分に満たされる。

[2]http://www.alice.columbia.edu/

4.1 考察

ユーザの興味は連鎖的に喚起されてゆくことが多い。これは、ユーザが抱いていた曖昧で不十分な興味が回答を読むに従い具体化され、新たな問題点に目が向くようになっていくためである。文書ネットワークではそのような連想的に沸いてくるユーザの興味を満たすことができるため、ユーザに深い理解をもたらすことが可能となる。

5 まとめ

本稿ではユーザの興味を網羅的に反映した文書ネットワークを用いて、ユーザの興味を満足する新しい検索手法を提案した。今後は実験による評価を進める予定である。

参考文献

[1] Burke, R., Hammond, K., et. al.: Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System, *Univ. of Chicago, Dept. of Computer Science Technical Report TR-97-05*, 1997.

[2] 松村真宏、大澤幸生、谷内田正彦：AAS：文書の組み合わせによってユーザの興味を満足する検索システム、人工知能学会誌 Vol. 14, No. 6, pp. 1177 – 1185, 1999.

[3] 山田誠二、大澤幸生：WWW における概念理解のためのナビゲーションプランニング、人工知能学会誌 Vol.14, No.6, pp.1125–1133, 1999.

[4] Salton, G. and Buckey, C.: Term-Weighting Approach in Automatic Text Retrieval, *Reading in Information Retrieval*, pp.323–328, 1998.