1S-5

作成文書のコンテクストを利用した検索クエリ生成システム

渡辺　一樹†　　東　基衛†

早稲田大学大学院　理工学研究科　経営システム工学専攻†

１．はじめに

近年のインターネットの普及に伴い、情報の発信や獲得が、誰でも容易に行えるようになった。そのような状況の中、企業の取り組みとして、各個人の成果物を組織全体で共有し、組織の知的生産性を向上させようとするナレッジマネジメントが注目されている。しかし知識の有効活用が行えず、失敗例が多いのが現状である。

そこで本研究ではナレッジマネジメントの一手法として、組織の知識を有効に活用した情報検索システムを提案する。

２．現状分析と問題点

組織のナレッジマネジメントが失敗に終わる要因として、以下の２つを挙げた。

１）各個人の情報収集活動が非効率である。

各個人が情報を収集する際、検索エンジンを利用した Web 検索を行う事が多い。Web 検索を効率的に行うためには、情報を絞り込むためのキーワードを、適切にかつ多数入力する必要がある。しかしユーザの知識不足により、抽象的なキーワードを少数しか入力できない。そのため、ユーザは膨大な検索結果の中から有用な情報を探さなくてはならない。

２）組織内の成果物を再利用するためのフレームワークができていない。

企業では、各個人により作成された文書を共有フォルダに蓄積する事が多い。しかし、その文書の再利用を促すようなフレームワークがなく、知識が埋もれてしまう場合が多い。また文書に関してもタイトルや作者しか把握できず、その文書がどのようなコンテクストを持っているかが把握しにくい。

３．研究目的と研究アプローチ

本研究はナレッジマネジメントを有用にする手法の一例として、組織内の文書から各個人の情報獲得活動に有効な情報を抽出し、利用する検索システムを提案する。本研究のアプローチは以下の２つとする。

1)ユーザの入力を基に、システムがユーザの検索クエリの想起の支援を行う。

現状の検索システムでは、ユーザの入力が適切でなければ、目的の情報に辿り着かない。

そこで本研究では、ユーザの入力を基にシステムが検索意図に則したクエリを複数提示する事で、ユーザの柔軟な想起を可能にする。

2)文書内容だけでなく、その文書のコンテクストを反映させたモデリングを実現する。

文書のタイトルやその内容だけでは、その文書のコンテクストを理解するのは困難である。

本研究では作成した文書が「どのような知識で作成されたか」をコンテクストと定義する。作成した文書だけでなく、その関連文書も利用してモデリングを行う事で、コンテクストを考慮した検索システムを実現する。

４．提案システム

本システムは組織内の文書を利用して、各個人の情報検索に有効な検索クエリを生成し、提供するシステムである。

本システムは文書登録と検索クエリ生成に分ける事ができる。文書登録においては、その文書の内容だけでなく、その関連文書（引用文書・同一作成者文書）を利用してコンテクストを抽出し、モデリングに反映させる。

検索クエリ生成の概要を図１に示す。

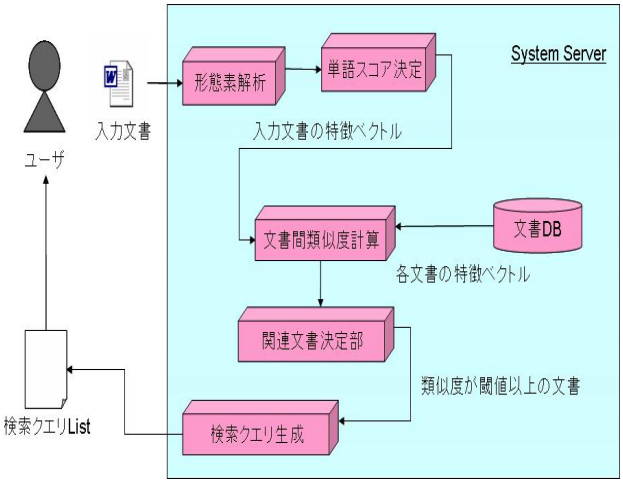

図１.検索クエリシステムの概要

ユーザはキーワードの代わりに文書を入力する。入力文書に対して、組織内にある全文書の中で、類似度の高い文書を抽出し、検索クエリ生成の際に利用する。システムは複数の類似文書から生成されたクエリをユーザに提示する。

4.1 文書からのコンテクスト抽出

文書内容から得られる情報のみでは、その文

A query generation System for Web Search using the Document Contexts
Kazuki Watanabe, Motoei Azuma
Dept. of IMSE, Graduate School of Sci. & Eng., Waseda Univ.

書が作成されたコンテクストを抽出するのは難しい。作成した文書はその内容だけでなく、文書が作成されるプロセスから抽出された情報も重要であると言える[1]。

そこで文書の特徴付けを行う際に、関連文書からその文書を捉えるようにする。以下にそのイメージを示す。

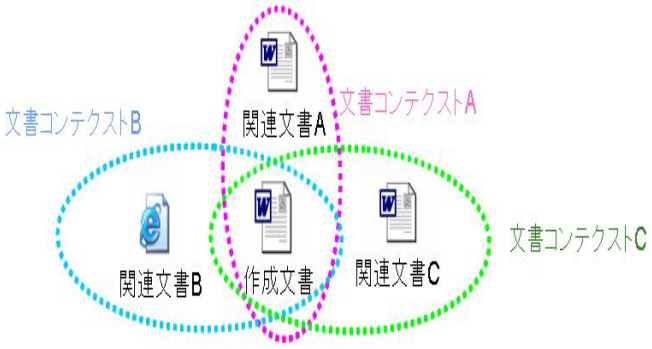

図 2.文書コンテクスト抽出のイメージ

1 つの文書を複数の関連文書で捉える事で、複数のコンテクスト表現が可能になる。

4.2 文書登録時の特徴付け

本研究では、組織内で作成された文書を登録する際、文書の特徴付けを行う。文書が作成されたプロセスを文書の特徴付けに反映するために、本研究では登録する文書の語だけでなく関連文書の語の重みも考慮する。

関連文書の語の重みを付与する事で、文書を作成したコンテクストを考慮に入れた特徴付けを実現する。登録する文書の特徴付けのアルゴリズムは以下の通りである。

＜文書の特徴付けアルゴリズム＞

①登録文書の各語の TF(Term Frequency)値を算出し、ベクトルを生成する。
②登録する文書に対する関連文書を取得する。
③式(１)により登録文書のベクトルを更新する。登録文書の単語の TF 値に登録文書と関連文書の TF 値の差分を加算する（登録文書と関連文書のコサイン類似度 sim により加算する TF 値を重み付ける。）

$$\vec{W}_{regist} = \vec{W}_{regist} + \alpha \sum_{k=i}^{N} sim(\vec{W}_{regist}, \vec{W}_{relationk}) \cdot \left| \vec{W}_{regist} - \vec{W}_{relationk} \right| \quad \cdots(1)$$

$\vec{w}_{regist}$:登録文書の特徴ベクトル

$\vec{w}_{relation:i}$:関連文書の特徴ベクトル

α ：重み

本研究では、2 つの類似文書間において、重要度の差が大きい単語を重要と考える。登録文書に関連文書との重要度の差分を加えることで、文書作成というプロセスにおいて重要であろう語の重みを大きくすることを考えている。

4.3 検索クエリの生成

従来の検索クエリ生成の支援法としては、文書から TF 値が高い語や、その語に共起する語を抽出して提示する手法が多い。しかしこの手法では、TF 値が非常に大きい語や抽象的な語の影響を受けやすく、ユーザが容易に思い浮かべる事の出来る語を多く抽出してしまうという問題点がある。

そこで本研究では、検索ユーザが入力した文書と、それに類似した組織内の文書の TF 値を算出し、その差分を単語のスコアとして算出する。これにより、低頻出であってもその文書のコンテクストを表現できるような語を抽出し、ユーザの検索クエリ想起を促す事ができると考える。検索クエリ生成のアルゴリズムを以下に示す。

＜検索クエリ生成アルゴリズム＞

①キーワードの代わりに文書（word 文書・PDF 文書）を入力する。
②入力した文書から TF 値の上位 N 語の単語を抽出し、ベクトル$\vec{W}_{input}$を生成する。
③組織内の全文書から各文書 TF 値の上位 N 個の単語を抽出し、ベクトルを生成する。
④組織内文書と入力文書の類似度を計算し、閾値を超えた文書を類似度の降順にソートする。
⑤閾値を超えた文書のベクトルを$\vec{W}_{similar}$とし、入力文書の TF 値の差分を式(2)より計算し、その値が大きい単語からリスト表示する。

$$\Delta w = \left| \vec{W}_{input} - \vec{W}_{similar} \right| \quad (2)$$

５．評価実験と考察

本研究の検索クエリ生成システムのプロトタイプを実装し、評価実験を行った。本システムによって生成された検索クエリを使って、Google scholar[2]における検索の適合率を計測したところ、高い適合率が得られた。

６．おわりに

本研究ではナレッジマネジメントの一手法として、組織内の文書を利用した検索クエリ生成システムを提案した。1 つの文書に関して、その関連文書を利用してコンテクストを抽出する事で、個人の検索に有用な検索クエリの生成が可能になった。今後さらなる検索精度の向上を図るため、キーワード抽出法などの改良を考える必要がある。

参考文献

[1]梅木秀雄他：”ナレッジワーク支援システム Trino の構想”.情報処理学会研究報告 2005-GN-55
[2]Google Scholar http://scholar.google.com/