質問文からの検索条件と提示情報の決定

伊吹 潤, 西野 文人
ibuki.jun,Nishino@jp.fujitsu.com
富士通研究所ドキュメント処理研究部

1　はじめに

我々の開発中の情報ワーク支援システム [1] では XML 検索の枠組を利用することによって高精度の検索と情報の加工のし易さの実現を図っている．本論文では全体の構成の中で, 新聞記事内容についての質問文を解析する部分と検索条件の生成, 実行を行なう部分の実装と性能評価を行なう．なおシステム全体の構成については [1] で別個に述べることとする．

2　今回の評価の対象とした分野

我々が今回の評価実験の対象としたのは, 記事に対して会社の活動全般とその対象となる製品や事業内容に関するマークアップを行なった日経新聞 90 年 - 97 年の 8 年分の XML 化文書データベースである．

我々は質問文の解析部の実装に当たって質問文をアンケートによって収集し、その中から 20 文を選択して、解析規則の開発等を行なった（表 1 参照）。

3　質問文解析・検索式生成の実装

本論文で扱うのは、システム全体の中でユーザからの文章による質問文を受け付けてその意図を解釈して検索を行なう部分である. この部分の実装に当たって, 我々は質問文解析部, 検索条件設定部, 検索実行部の 3 つの要素による構成をとった. 以下、各要素毎に説明する.

3.1　質問文解析部

質問文解析部では外部に解析用規則のテーブルを持ち, それを参照することで処理を行なう. 処理内容は、次のような段階に沿って行なう.

1. 質問文の文パターンの認定

   解析規則中に登録された約 30 のパターンとの照合によって質問文の文型の認識を行なう.

2. 質問詳細（条件指定, トピック項目）の解析

   質問文中の製品や組織等の記述を解析して、検索条件と直接対応する形でタグ付けを行なう. 又, 質問文のトピックとなる項目をマークアップする.

実際の解析の例を下に示す.

入力文：幼児向けのパソコン関連商品にはどのようなものがあるか

解析結果：

```
<検索要求 q="製品 ">
  <製品条件部>
    <キーワード> 幼児 </ キーワード> 向けの
    <キーワード> パソコン </ キーワード> 関連
  </ 製品条件部><製品クラス> 商品 </ 製品クラス>
  <要求部> にはどのようなものがあるか </ 要求部>
</ 検索要求>
```

「製品条件部」に示されているのは検索条件の設定に必要な情報である. 又「q="製品 "」の部分は質問文のトピックが製品にあることを示す.

3.2　検索条件設定部

ここでは検索条件の拡張と検索の際の XML タグの同定の 2 つの処理によってテキスト検索時の検索条件を生成する.

1. 質問拡張（同義語への条件拡張）

   システムは各タグ項目毎に同義語辞書をもっており, これらを参照することによって同義語による条件の拡張を行なう.

文例
キーボードを販売している会社を知りたい.
幼児向けのパソコン関連商品にはどのようなものがあるか
冷凍のすしを開発したメーカは？
木でできたマウスというのはありますか？
生ごみを処理する装置について教えて下さい.
Ｏ１５７対策で注目されている商品にはどのようなものがありますか
日本語を中国語に変換するための辞書について知りたい.
ぶらっとホームという会社はどんなものを売っているのか.
ＯＣＲソフトを開発している会社にはどのような会社があるか.
ズーという会社はどのような活動をしているのか.
ＮＨＫが放送した番組の映像を使ったＣＤ－ＲＯＭソフトにはどのようなものがあるか
バンダイがゲームボーイ用に開発したソフトとは？
ビジネスマンの教育研修を目的としたソフトを開発している会社とそのソフトについて知りたい
Ｍａｃの上で動く占いソフトを知りたい
フラッシュメモリーを使った音声を記録する装置を開発したのは
電車の運転のシミュレーションソフトではどんなものがあるか.
実売価格が 13 万円以下のパソコンにはどのようなものがあるか
ＭａｃＯＳっていつから発売
電話帳のデータを使ったソフトとしてはどのようなものが提案されていますか
松本にある医療関係のソフトを開発している会社って

表 1: 質問文

2. XML タグの同定

テキストベース中のタグ項目に対する条件を決定する．各記事には開発, 販売, 提携等の記事分類が付加されており、ここでは 1) 記事分類の決定 2) 各記事分類に付随する下部タグ項目の決定という 2 段階での処理を行なう.

3.3 検索実行部での処理

ここでは検索条件を利用して実際に検索を行なうが、検索条件をそのまま利用するのではなく検索結果を睨みながら必要に応じて調整した後に実行している．このような構成をとったのは以下の理由による.

XML 文書検索における基本的問題

XML 検索においてはタグ項目名と値の両方が一致しなければいけないためキーワードによる検索に比べ制限が厳しく、頻繁に検索に失敗し検索結果が空になってしまう．その要因としては次の 2 つが大きく影響している。

- 製品の仕様に関する記述は発表や開発のイベントを表す文中だけでなく、記事中の他の文にも分散して記述されていることが多い. これらの記述は, 製品情報としてタグ付けできていないために検索の際に除外されてしまう.
- 現在のタグ体系では, 同一の製品に関する情報でも、開発の段階によって異なる種類の記事（開発, 販売等）となってしまう．特に製品情報を求める場合, 記事種の指定を行なうことによって限定された情報しか得られないことがある.

ＱＡシステムの使用目的（基礎的な調査等）を考えるとなんらかの検索結果を得ることが検索ゴミの混入よりも優先すると判断し、このために検索失敗時に検索条件緩和操作を導入し, 必要に応じて検索条件を緩めることとした。

検索条件の緩和操作

我々が導入した緩和操作の種類について説明する.

- 製品に関する検索条件の緩和

  我々の製品検索条件に対しての態度は次の通りである．基本的な製品のカテゴリについては XML 検索の精度を保証するためタグ条件から動かすことはない。他のキーワードに対しては制限がきつ過ぎる場合はキーワード検索と同じ条件まで制限を緩和することによって空でない検索結果を確保する.

このような目的のために製品／開発内容を指定するキーワード群を重要度順にソートし、重要度の最も低いものから順に以下の条件緩和操作を行なう.

1. タグ指定のない一般のキーワード条件として対象文字列を解釈する.
2. 上記の場合でも検索に失敗する場合は、条件自体を削除する.

- 関連記事種への条件緩和

  市販されている製品に対する検索要求があれば、システムは記事種を販売情報として検索を行なうが、もし市販された製品に関する情報がなければ、研究段階の装置等の情報を提示するのが望ましいだろう。このために我々は関連記事種同士のリンク情報を用意し、緩和操作として関連記事種への記事種の変更を行なうこととした.

  ex.　開発情報　←→　販売情報

なお、現状では緩和操作は救済措置として扱い, 検索が失敗に終った時（検索結果が空）のみに適用するようにしている.

4　検索性能の評価

4.1　キーワード検索との全般的な性能の比較

検索性能の比較のため、我々は文章をそのまま入力してキーワードを抽出して tf・idf に基づいてランキングを行なう検索システム（以後キーワード検索と呼ぶ）を用意し, 同一の質問文を入力して検索結果の比較を行なった. キーワード検索においては上位5位までを検索集合とした場合の適合率（適合率1）を出した.XML 検索の場合は適合率（適合率2）と再現率を計算している. 再現率の評価の際にはキーワード検索の結果から人手で選択した正解ファイルを作成して利用した.

適合率1	適合率2	再現率
0.35(34/95)	0.81(82/101)	0.61(27/44)

表 2: キーワード検索との比較

XML 検索とキーワード検索の場合の検索結果の大きさはほぼ同じであるが、適合率はキーワード検索の場合の 35% に比べ、XML 検索の場合は 81% であり、40 ポイント以上の差がある.
個別の文例をみると XML 検索の結果が大きいのは、特に単語の意味にあいまいさがあるような場合である。タグ名と共に条件として指定することによって意味を限定することができるからである.

ex.　松本にある医療関係のソフトを開発している会社って？
→所在地が松本と指定する（人名とは混同しない）

次に XML 検索の再現率を評価してみた結果は 61% となった. ここでの 16 件のもれの原因は構造解析の失敗が 11 件、質問文との柔軟なマッチングが必要なものが 5 件となっている。

4.2　条件緩和操作の効果

次にシステムの検索性能が導入した条件緩和操作の改善効果を評価した結果を示す.

	適合率	再現率
緩和操作なし	0.98(64/65)	0.38(17/44)
自動緩和操作	0.81(82/101)	0.61(27/44)
手動緩和操作	0.81(86/106)	-

表 3: 緩和操作の寄与

適合率の低下（98% から 81% へ）と引き換えに再現率の向上（38% から 61%）を得ている。人手でタグ条件を変化させた場合に比べて性能はわずかに劣るが、充分な効果を得ていると考える。

ここで緩和操作が有効に働いた条件の例を見るとソフトのプラットホーム, 価格, 用途等に対応する部分であることが分かる。これらの大部分は製品の発表文とは別の文に書かれており、情報抽出の対象から洩れることが多いため, タグ項目条件では検索もれとなってしまっていた。

ex.　Mac の上で動く占いソフトを知りたい
初期条件→ <製品種>Mac</>
条件緩和 → <keyword>Mac</>

関連イベントへのタグ緩和操作は現状では緩和操作が救済措置とされているために実際に働いたのは次の場合のみである. ここでは開発したことが陽に書かれていないために開発記事を対象にした最初の検索に失敗している。

ex.　バンダイがゲームボーイ用に開発したソフトは？

記事の文面：
バンダイは七月にも、任天堂の携帯型ゲーム機「ゲームボーイ」
と専用ソフト「ゲームで発見！たまごっち」をセット販売する。

4.3　ブラインドテストの結果

最後に新たに8つの質問文を設定し、検索部でのチューニングを全く行なわない状況（質問文の解析部での単語登録のみ）での評価を行なった結果を示す．ここでの適合率1，適合率2，再現率の計算方法は4.1におけるものと同様である．ただし本実験においては緩和操作の結果，検索件数が100件を越えた場合には検索失敗として検索を打ち切った．

適合率1	適合率2	再現率
0.39(31/80)	0.78(35/45)	0.25(4/16)

表4: ブラインドテスト（8文）の結果

ブラインドテストの場合でもXML検索の適合率は78%とキーワード検索の39%の2倍程度となり、適合率のメリットが依然大きいことがわかる。ただし再現率は25%と非常に悪い。もれの原因としては文書登録時の解析失敗が一番大きく、発表述語のない製品の説明記事や述語の表現のパターンが充分解析されていないのが原因となっている。

5　まとめ

XML検索の性能について　XML検索の適合率の向上に対する効果は明らかであり，タグ情報による表表示の効果も相まってユーザにとって見やすい情報を提示できていると言える。

キーワード検索の場合も検索件数が大きい場合はランキングなどを利用した絞り込みが不可欠であり、同程度の検索集合を得る場合XML検索の方が優れていることは確かだろう。

改善項目　ただし絶対的な性能として検索における再現率の向上が今後の目標のひとつであることも事実である。現在の対象分野での性能の改善のための方策として考えることを以下に示す．

- 浅いレベルでの解析の導入

  特にブラインドテストでは検索もれの原因としてテキスト解析の失敗が多く，テキスト解析能力の向上が急務である．そのためには，対象とするテキストの解析パターンの充実と共に，主語，目的語等の文全体の構文解析情報，日付等のローカルな解析によって抽出できるエンティティの情報をカバーし，意味的な解析に失敗した場合にも何らかの情報が得られるようにすることが必要だと考えている．

- 製品記述の精密化

  ブラインドテストでは製品条件緩和の結果，検索件数が700件を越え，検索が失敗した場合もある．全体の評価では数値的な影響は少なかったが将来は製品記述の解析をより細かな項目毎に行ない，テキストの登録や条件緩和操作の対象の選択に活用すべきだろう．

終りに　検索結果の評価は，実際には数値以上に，最終的な利用目的に役立ったかで判断する必要がある．

例えば製品の主なメーカーを知りたい場合は，同一のメーカーの記事は一件存在すれば充分となり、検索結果はメーカー数がどの程度検索できたかで評価することになる．一方で検索対象となる新聞記事のデータベースに情報がなければ複数のデータベースを利用することも必要だろう。そのためにもまず、システムの試用を通じて実際にユーザの評価を収集していきたいと考えている．

参考文献

[1] 西野文人，伊吹潤：“質問文からの検索条件と提示情報の決定”，言語処理学会第6回年次大会 (2001)