1X-7

類似文書検索における 高精度レリバンスフィードバック方式の開発

稲場靖彦　多田勝己　菅谷奈津子　松林忠孝
(株)日立製作所 ビジネスソリューション事業部

1. はじめに

近年、業務の質の向上や効率化のために、組織内の知識共有や再利用を実現する知識管理システムに対する要求が高まっている。

組織内に大量に蓄えられた知識の中から、必要なものを取得するための手段として、類似文書検索技術がある[1]。これは、ユーザが検索条件として入力した文章（以下、種文書と呼ぶ）と内容（概念）が似た文書を検索する技術である。

この類似文書検索の精度を向上させる技術として、レリバンスフィードバックがある[2]。この技術を用いると、検索結果として提示された文書に対してユーザが評価を与え、これをフィードバックして対話的に再検索を繰り返し、より目的とする文書を入手することができるようになる。

しかし、検索結果に対する「適(適切なもの)」という評価は的確にフィードバックできても、「不適(不適切なもの)」という評価を的確にフィードバックする手段がなかった。これは、「不適」と評価された文書から、ユーザにとって不要な概念だけを抽出してフィードバックする仕組みがなかったためである。そのため、「不適」という評価を単純にフィードバックすると、所望とする文書までが再検索結果から除外されてしまい、再現率が低下してしまうことになる。

本研究は、上記問題を解決するため、ユーザの「適」「不適」の評価を的確に検索条件へ反映させるレリバンスフィードバック方式を開発したものである。

2. 従来方式の問題点

一般に類似文書検索では、種文書からその内容を特徴づける語（以下、特徴タームと呼ぶ）を抽出して、重要度に応じた「重み」づけを行なう。そして、重みが大きい特徴タームを多く含む文書ほど高い類似度(スコア)がつくように、データベース内の各文書にスコアを付け、スコア順にユーザに提示する。

従来、レリバンスフィードバックとしては、以下のような処理方式が提案されている[2]。

(1) ユーザが「適」と判定した文書から特徴タームを抽出し、その特徴タームの重みを積算する。
(2) ユーザが「不適」と判定した文書から特徴タームを抽出し、その特徴タームの重みを上記積算値から減算する。
(3) 上記 (1)(2)によって修正された重みを用いて、スコア算出を再び行う。

しかしながら、検索結果として提示される文書には、内容が所望のものでない場合でも、ユーザが所望する概念を表す特徴タームが含まれている場合が多い。したがって(2)の処理を単純に行なってしまうと、その所望の概念を表す特徴タームの重みを減算してしまうことになる。その結果、所望する文書のスコアまで下がり、検索結果として提示されなくなってしまう恐れがある（図1）。

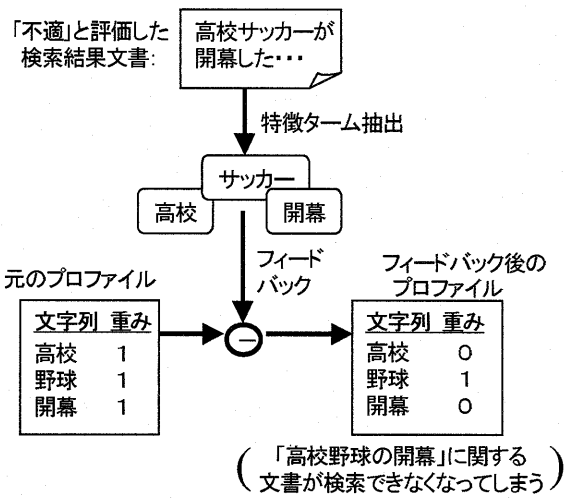

図1.「不適」フィードバックの問題点

そこで本研究では上記の問題に鑑み、「適」という評価のみならず、「不適」という評価も的確にフィードバックできる方式の開発を課題とした。

Development of Highly Precise Relevance Feedback Method for Relevant Document Search.
Yasuhiko INABA, Katsumi TADA, Natsuko SUGAYA and Tadataka MATSUBAYASHI
Business Solution Systems Division, Hitachi, Ltd.

3. 高精度レリバンスフィードバック方式の開発

種文書や、ユーザが評価を与えた文書から抽出される特徴タームには、以下の種類がある (図 2)。

A： 種文書および「適」と判定した文書だけから抽出されるもの

B： 種文書および「適」と判定した文書から抽出され、かつ「不適」と判定した文書からも抽出されるもの

C：「不適」と判定した文書からだけ抽出されるもの

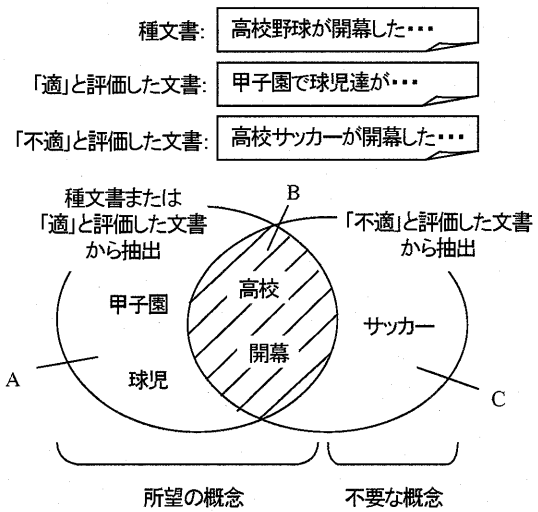

図2. 文書から抽出される特徴タームの種類

従来方式では、検索結果に対するユーザの「不適」という評価をフィードバックする際に、B に相当する特徴タームの重みも減算してしまうため、ユーザの所望する概念を表すような特徴タームの重みまでが減算されてしまう。

しかし本来はC に相当する特徴タームだけを、ユーザの不要な概念を表しているものであると判断する必要がある。

したがって本研究における方式では、「不適」という評価をフィードバックする際にBの特徴タームの重みは減算せずに、C の特徴タームの重みを減算する、すなわち「負の重み」を与えることとした。すなわち、下記方式によるフィードバックを実現する。

> 「不適」と評価された文書から抽出される特徴タームのうち、「適」と評価された文書から抽出される特徴タームを除いて負の重みを与える

これにより、所望しない概念を表すと考えられるC の特徴タームが出現する文書についてだけスコアを下げることができるため、ユーザの「適」「不適」の評価を的確に反映したレリバンスフィードバックが実現できるようになる。

4. 検索精度の評価

本レリバンスフィードバック機能を類似文書検索システムに実装し、被験者6人による2テーマずつの実験を行ない検索精度を評価した(検索対象データベース：新聞記事 1 ヶ月分(約 12,000 件))。実験の手順は以下のとおりである。

① 種文書を設定して類似文書検索を行う
② 上記①の結果に対して「適」の評価を3回行う
③ 上記②の結果に対して「不適」の評価を3回行う

この実験の結果、「不適」という評価を行なっても再現率を低下させることなく適合率を上昇させることができることが分かった(図 3)。

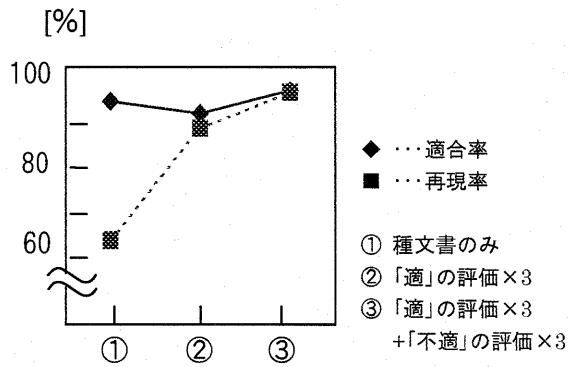

図 3. フィードバックによる適合率・再現率の推移

5. まとめ

ユーザの「不適」という評価を的確に反映することができるレリバンスフィードバック方式を開発した。評価の結果、「不適」という評価によっても再現率を低下させることなく適合率を向上できることがわかった。

参考文献

[1] 松林他：「知識指向文書管理基盤の開発(5) n-gram 方式に基づく概念検索」、情報処理学会第59 回全国大会 3-145

[2] William B.Frakes 他 ： 「 Information Retrieval」 pp241～263, Prentice Hall