社内文書検索システム（6）-未出優先ランキング-

原 雅樹
NEC サービスプラットフォーム研究所

1. はじめに

大量の電子化文書が格納される企業内ポータルにおいて、検索機能はユーザが所望の文書を探し出すために不可欠な機能である。我々は、文書検索システムにおけるランキング機能の一手法として、未出優先ランキング手法を提案する。従来のランキング手法では、欲しい文書が見つからず検索クエリを変更して再検索した場合でも、同じ様な文書が検索結果上位に表示されるという課題があった。未出優先ランキング手法では、再検索の際に、ユーザに対して未だ提示していない文書を優先するようにランキングすることで、文書を見つけ易くできる。

2. 従来のランキング手法と課題

ランキング手法の分類

従来のランキング手法は、文書属性によるランキング、検索クエリとの適合度によるランキング、ユーザ嗜好によるランキングに大別できる。

文書属性によるランキング手法は、各文書に付与されたタイトル、著者、人気度、登録日などの属性情報を使って、文書をタイトル順、著者順、人気順、新着順などにランキングする。

検索クエリとの適合度によるランキング手法は、ユーザが入力した検索クエリと各文書との類似度を算出し、算出した類似度に従って文書をランキングする。

ユーザ嗜好によるランキング手法は、システムが保有するユーザの嗜好情報と、各文書との類似度を算出し、算出した類似度に従って文書をランキングする。

従来手法によるランキングの傾向

検索機能を使って所望の文書を見つけるためには、目的の文書がヒットしそうな検索クエリを予想し、入力する必要がある。しかし、検索クエリを適切に設定することは難しく[1]、多くの場合、ユーザは検索クエリを試行錯誤しながら所望の文書を探索することとなる。実際に使用された検索クエリを分析すると、検索クエリを『セキュリティ対策』、『セキュリティチェック』、『セキュリティソフト』など、類似した語句に言い換えて検索を繰り返すケースが多い。しかし、このように類似した検索クエリを設定した場合、再検索しても同じ文書が多く検索にヒットし、従来のランキング手法では、検索クエリ変更前と似通った検索結果（文書の一覧）が生成されることが多い。

新しいランキング手法の要件

ユーザが同じ検索目的を保持したまま検索クエリを変更する時、ユーザはそれまでの検索結果に含まれる文書に満足できず別の文書を見たいはずである。しかし、従来のランキング手法では検索クエリを変更したにもかかわらず、変更前と似た検索結果が表示され、その結果、ユーザが希望する文書を取得できずに検索を諦めてしまうという問題があった。

この問題を踏まえて我々は、同じ検索目的で再検索を繰り返した時に、欲しい文書を見つけやすくする新しいランキング手法の要件を定義した。新ランキング手法は、『同一目的で検索しているが、うまく検索できていない』*(発火要件)*というユーザ状況を検出し、この状況の時に、『一連の検索操作で未だ提示していない文書を優先してランキングする』*(ランキング要件)*ことと定義した。

3. 未出優先ランキング

新ランキング手法の発火要件、ランキング要件を踏まえ、未出優先ランキングを設計した（図 1）。未出優先ランキングでは、既存の検索エンジンでランキングされた検索結果をベースに、未だ表示していない文書を上位に表示するよう文書の並びを変更して検

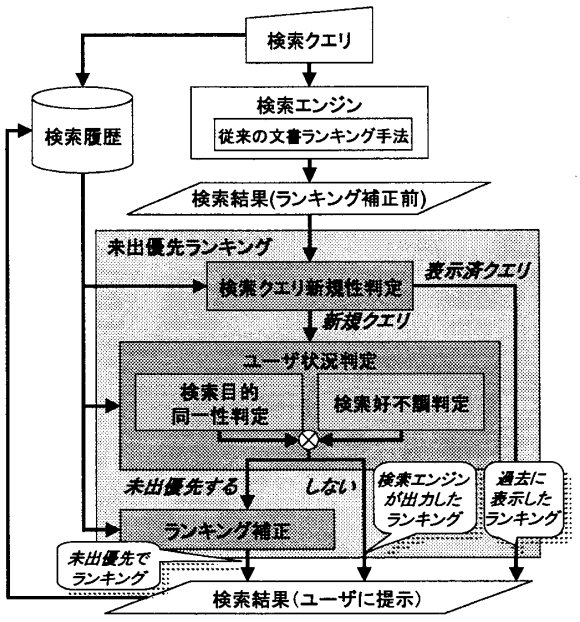

図 1 未出優先ランキングの構成

An In-house Documents Retrieval System (6) –A ranking method giving priority to unseen documents-
Masaki HARA
Service Platforms Research Laboratories, NEC Corporation

索結果を生成する。以下、各モジュールの動作について説明する。

3.1. ユーザ状況判定

ユーザ状況判定モジュールでは、ユーザの状況が発火要件を満たすか否かの判定を行う。

ここで、発火要件をユーザが『同一目的で検索している』と、『うまく検索できていない』の 2 つに分割し、それぞれについて、検索目的同一性判定、検索好不調判定の 2 つのサブモジュールにより、未出優先ランキングが必要な状況かどうか判定する。

3.1.1. 検索目的同一性判定

検索クエリによりヒットする文書集合を使って、検索目的が同一であるかどうかを判定する(図 2)。

ユーザがこれまでに同一の検索目的で n-1 個の検索クエリ $Q_i(1 \leqq i \leqq n-1)$を入力し、今回新たに n 番目の検索クエリ Q_n を入力したとする。また検索クエリ Q_i により検索される文書集合を S_i とする。このとき、検索目的同一性を示す指標(目的同一性指標 I_n)を、式(1)により算出する。

$$I_n = \max\left(\frac{|V_n|}{|U_{n-1}|}, \frac{|V_n|}{|S_n|}\right) \quad \cdots(1)$$

$$U_{n-1} = \bigcup_{1 \leq i \leq n-1} S_i, \quad V_n = U_{n-1} \cap S_n$$

式(1)で算出する目的同一性指標 In は、前回(n-1 回目)までに検索された文書集合の和集合 U_{n-1} と、今回(n 回目)の検索での文書集合を S_n との重なり具合を表す。

検索目的同一性判定では、U_{n-1} と S_n とが 50%以上重なった状態の時に、検索クエリ Q_n は、それまでの検索クエリと同一目的であると判定することとし、目的同一性指標 I_n が閾値 0.5 以上のとき、検索クエリ Q_n は、ランキング補正の必要ありと判定する。一方、I_n が閾値 0.5 未満の場合は、検索クエリ Q_n は異なる検索目的で入力されたと推測し、ランキング補正の必要なしと判定する。

3.1.2. 検索好不調判定

うまく検索ができない状況として、『検索を繰り返しても選択したいと思う文書が見つからない』検索シーンを設定し、検索の好不調を判定する。

検索の好不調を示す指標(好不調指標 C)を、総操作回数 No、検索クエリ入力回数 Ns、検索結果での文書選択回数 Nc を用いて、式(2)により算出する。

$$C = (Nc + 20 \times e^{-1 \times No}) / (Ns + 1) \quad \cdots(2)$$

式(2)で算出する好不調指標 C は、検索クエリを入力した回数 Ns に比べて、文書選択回数 Nc が小さい時に小さな値となり、検索が不調であると判定する。(なお、分子の第 2 項は操作回数 No が 2 回未満の時に検索不調と判定しないための調整項である)。

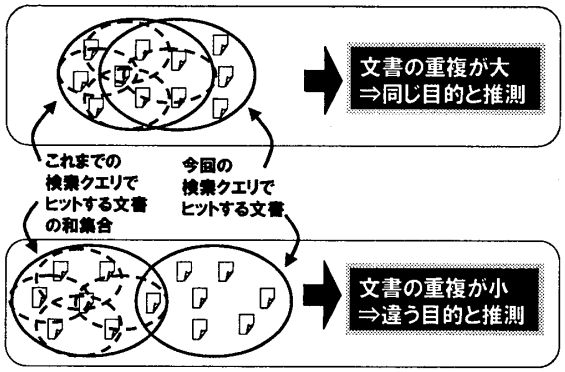

図 2 検索目的同一性判定

検索好不調判定では、1 回の検索クエリ入力につき最低 1 つの文書が発見できなければ、検索は不調であると判定することとし、好不調指数 C が閾値 1.0 未満のとき、ユーザの検索行動は不調であると推測し、ランキング補正の必要があると判定する。

3.2. ランキング補正

ユーザ状況判定モジュールにて、未出優先ランキングを行うと判断した時に、ランキング補正を行う。

文書のこれまでの表示回数 Nd、選択回数 Nc、現在の検索クエリで検索した際に検索エンジンで付与されたランキング値(スコア)R としたときに、補正ランキング値 R'は、式(3)により算出する。

$$R' = R / (Nd - Nc + 1) \quad \cdots(3)$$

各文書について、式(3)で補正ランキング値を求め、補正ランキング値でソートすることで、未出優先ランキングによる検索結果を生成する。

3.3. 検索クエリ新規性判定

検索履歴を参照して、入力された検索クエリがユーザにより過去に使用されたかどうかを判定する。これは、同じ検索クエリで再検索した際に表示される検索結果の内容が変わると、ユーザが検索サービスに不信感を抱くことから、過去に使用された検索クエリであれば、以前に表示した検索結果を出力する。新規の検索クエリであればユーザ状況判定、ランキング補正処理を行って、ランキングを変更する。

4. おわりに

本稿では、ユーザが同じ検索意図で再検索を繰り返している時に未だ表示していない文書を優先的に表示する未出優先ランキング手法について述べた。今後、プロトタイプエンジンによるユーザ評価実験を通して、有効性を検証する予定である。

参考文献

[1] White,R.W.,et.a.:Supporting exploratory search, Comm. ACM, Vol.49, No.4, pp.36-39 (2006)