

ディレクトリ構造に着目した企業内文書向け ランキング方式の検討

四ッ谷 雅輝[†] 松林 忠孝[†] 弥生 隆明[†]
野田 十悟[†] 吉田 豊[†]

検索者が自分の検索要求に適合した文書を探し出すためには、検索システムによって文書の内容を的確に評価できることが重要である。文書の内容を評価する方法としては、文書に含まれる検索語の出現数に基づいて評価する方法の他にも、文書間に張られたハイパーリンクを解析して、文書間にある関係から文書の内容を評価する方法などがある。本研究では、検索者に必要とされる文書の多くが、人手によってあらかじめ分類されたディレクトリに格納されているという傾向に着目し、ディレクトリの解析結果から文書の内容を評価する指標を提案する。本手法は、ディレクトリ構造に着目するものであるため、共有ファイルサーバ上の文書の検索にも適用可能である。

Examination of Ranking Method that Focuses on Directory Structure for Enterprise Document Search

Masaki Yotsutani[†] Tadataka Matsubayashi[†] Takaaki Yayoi[†]
Jugo Noda[†] Yutaka Yoshida[†]

In order to search documents relevant to user's information need, it is important that a search system can evaluate the contents of documents appropriately. As the method of evaluating the content of the document, there is a method based on the number of appearance of retrieval words in the document. As other methods, there is also a method based on the relation a document has hyperlink each other. In this paper, we propose a method for evaluating the contents of documents from an analytical result of the directory, focusing on the fact that many documents tend to be stored in the directory that classified by judging the contents of them. We can also apply our proposed method to search documents on shared file servers since our idea is based on directory structures.

1. はじめに

近年、企業内に蓄積された膨大な文書から検索者の所望する情報を探し出す企業内文書検索の重要性がますます高まってきている。これを受けて、企業内文書検索に求められる検索技術に関する研究が盛んに行われている。

検索者の要望を満たす検索システムを検討するためには、文書検索を行う上で必要となる

ステップを検索者の視点から整理することは重要である。検索システムを利用して、文書検索を行う上で必要となるステップは、大きく次の3つのステップに分けて考えることができる。3つのステップとは、(1)所望する情報を探すための検索条件を考え、検索システムに入力するステップ、(2) 検索条件として入力された文字列(以下、検索タームと呼ぶ)を含む文書(以下、ヒット文書と呼ぶ)を取得するステップ、(3)

[†] (株)日立製作所 ソフトウェア事業部
Hitachi, Ltd. Software Division

取得したヒット文書の中から所望する情報が含まれる文書(以下、目的文書と呼ぶ)を探し出すステップ、である。これら3つのステップを経て、検索者は文書データベースから目的文書を取得することができる。我々は、これまで検索者が容易に目的文書を取得できるように、(1)のステップに対しては、検索条件の入力の負担を軽減するために、同義語や異表記を自動的に展開する技術の開発、(2)のステップに対しては漏れなく高速に検索する技術の開発を行ってきた。

しかしながら、近年、電子化文書の数の増大に伴い、検索結果として得られるヒット文書の数も増大している。その結果、ヒット文書から目的文書を探し出す(3)のステップが、検索者にとって大きな負担になってきている。このため、ヒット文書から目的文書を探し出す検索者の負担を軽減する技術が求められている。ヒット文書の中から目的文書を探し出す負担を軽減する技術として代表的な例を以下に示す。

(a) ランキング技術

想定した目的文書の特徴に基づき、ヒット文書のスコアを算出し、スコアの降順にヒット文書を表示する技術である。本技術により、検索結果の上位には、スコアの高いヒット文書が表示されるため、目的文書を探し出す検索者の負担を軽減することができる。

(b) 要約表示技術

検索条件などに基づき、ヒット文書の要約を生成し、検索結果一覧にヒット文書の情報として表示する技術である。本技術により、所望する情報がヒット文書に含まれているか否かを検索結果一覧上で判断しやすくなるため、目的文書を探し出す検索者の負担を軽減することができる。

(c) クラスタリング技術

検索条件などに基づき、ヒット文書を内容に従って分類し、分類されたクラスごとに検索結果一覧に表示する技術である。本技術により、所望する情報が含まれているヒット文書が否かを分類されたクラスごとに判断することができるため、目的文書を探し出す検索者の負担を軽減することができる。

上記(a)～(c)の技術は、文書の内容に基づき、処理が行われるため、検索システムによって文書の内容を的確に評価できることが重要である。

文書の内容を評価する方法としては、文書に含まれる検索タームの出現数に基づき評価する方法の他にも、文書間に張られたハイパーリンクを解析して、文書間にある関係から文書の内容を評価する方法もある。本研究では、企業内に蓄積された文書の格納形態に着目し、ディレクトリの解析結果から文書の内容を評価する指標を検討した。

本研究では、閲覧すべきヒット文書の優先順位を提供するランキング技術に重点を置き、文書内容の評価指標をランキング技術に応用することを想定して検討を進めた。

本論文の構成は、次の通りである。2章では、本研究の関連研究について紹介する。3章では、企業内文書検索におけるランキング方式の課題について述べる。4章では、課題に対するアプローチについて検討し、課題解決の実現方法を提案する。5章では、提案方式の評価実験の結果を示し、本結果について考察する。6章では、本研究のまとめと今後の課題について述べる。

2. 関連研究

企業では、イントラネットの普及に伴い、Webサーバなどを用いて、従業員へ向けた情報発信がされるケースが多くなってきている。このような状況を受け、企業内のWebサーバで公開されている情報から、従業員が所望する情報を効率良く取得できるように、ハイパーリンクの解析結果を用いてランキングを行う検索システムを構築する動きがある[1]。

ハイパーリンクの解析結果を用いるランキング技術としては、Google¹社のPageRank方式[2][3]などがある。PageRank方式とは、Webページ間のリンク構造にランダムウォークモデルを適用し、WWW(World Wide Web)上に存在する全Webページへの遷移確率を基にスコアを算出する方式である。このスコアは、WWW上に存在する各Webページに対する被参照度とみなすことができる。このモデル化は、「有用な

¹ Googleは、Google Inc. の登録商標です。

Web ページは、多くの文書からハイパーリンクが張られている」という WWW における実モデルに合致する場合が多く、目的文書に高い評価を与えることができる。

しかし、企業内では、従業員が必要とする情報は、Web サーバの他にも、共有ファイルサーバやデータベースに含まれている可能性がある。事実、これらのデータソースを検索対象に含めた検索システムを望む従業員も多い[4]。このため、文書間に張られたハイパーリンクの存在が前提となるランキング方式では、企業内文書検索では、不十分であることが考えられる。これに対するアプローチとしては、PageRank 方式と、ハイパーリンクに依存しない評価指標によるランキング方式を併用する手法が提案されている[5]。これに用いられる文書の評価指標の例としては、検索タームの出現数や、更新日時の新しさをスコアに反映させるものなどが挙げられる。

3. 本研究の着眼点

前章で示したように、企業内文書検索では、複数のデータソースに対して検索を行いたいという要望が高まっている。そこで、本研究では、比較的小規模な企業でも、導入される可能性の高い Web サーバと共有ファイルサーバ上の文書を検索対象とした場合の検索に着目し、目的文書を検索結果の上位に表示できるランキング方式について検討した。

3.1. 現状の問題点

ランキング技術とは、想定した目的文書の特徴に基づき、ヒット文書のスコアを算出し、スコアの降順にヒット文書を表示する技術である。従って、検索結果のランキングが適切に行われない場合、ランキングアルゴリズムで想定された目的文書の特徴と適用先の検索対象に含まれる目的文書の特徴の間にギャップがあることが問題となる。

本研究では、企業内文書に対するランキング方式の検討を進めるにあたり、自社内の Web サーバを検索対象とした検索システムを試験的に構築し、試用調査を実施した。本システムでは、Web サーバと共有ファイルサーバ共に適用可能な TF ランキング方式を用いるものとした。なお、TF ランキング方式とは、検索タームを多く含む文書を検索タームに関連が深い文書

とみなし、検索結果の上位にランキングする方式である。

まず、本システムの利用ケースの設定にあたり、以前より運用されていた自社内の Web サーバを検索対象とした検索システムに入力された検索条件を分析した。この結果、検索条件として製品名が用いられる場合が多いことが判明した。これより、本システムの利用ケースの一例として、「製品情報の調査」を想定した。

「製品情報の調査」を利用ケースと想定した場合、検索者にとって、製品の価格、機能あるいは関連製品に関する情報など、多岐に渡る情報が必要と考えられる。そこで、試用調査では、検索タームとして製品名（以下、製品 A とする）を設定し、製品 A に関して様々な情報が記載されたページへ辿ることができる製品紹介トップページを目的文書と設定した。

設定した条件の下、試用調査を行った結果、TF ランキング方式による検索結果の最上位の文書は、「製品 A の拡販メールの配信履歴一覧」であった。試用調査では、検索者は製品 A の価格や機能など、多岐に渡る情報が必要であると想定したため、「拡販メールの配信履歴一覧」の文書は、検索者の目的を満たすものではないと判断した。

TF ランキング方式が企業内文書検索で有効に機能しない原因を分析するため、検索対象の調査を実施した。調査の結果、検索対象には、ホームページ、事務報告書、表データなど多様な内容の文書が混在しており、本方式では、ログデータなど特定の語が繰り返し用いられる文書を高く評価してしまうことが判明した。

以上より、企業内文書検索の検索対象には、多様な文書が混在するため、文書のテキストから、文書の内容を適切に評価するのは困難な場合があることが判明した。

3.2. 企業内文書向けランキング方式の課題

企業内文書を検索対象とした場合、検索者にとって必要な情報は、一文書に集約されて記載されているとは限らず、複数の文書に跨って記載されていることが考えられる。そこで、本研究では、個々の文書のみを評価するのではなく、同一トピックに関する情報が記載された文書の集合を抽出し、抽出された文書集合の評価も反映できるランキング方式を考案することを課題とした。

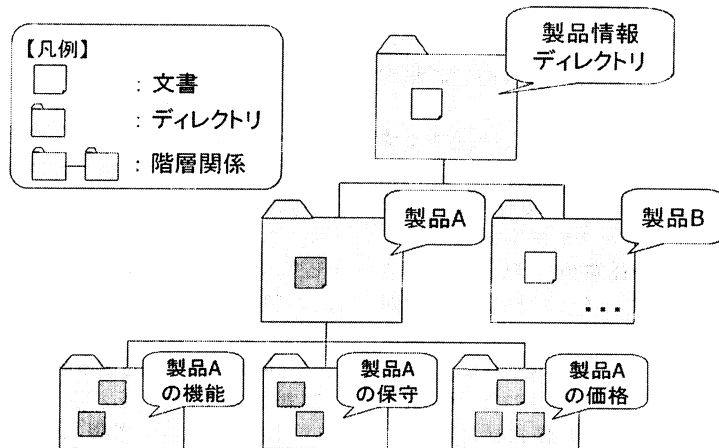


図1 製品Aに関する文書の格納形態の例

4. 方式検討

4.1. 課題に対するアプローチの検討

3.2節で示した課題を解決するためには、同一トピックに関する情報が記載された文書の集合を抽出する必要がある。本研究では、同一トピックに関する情報が記載された文書の集合を抽出する手法として、Web サーバおよび共有ファイルサーバ上の文書に共通する特徴である、ディレクトリによる格納形態に着目したアプローチを検討した。本研究では、アプローチの検討にあたり、企業内文書の格納形態を次のようにモデル化した。

- 企業内で蓄積される文書は、効率の良い情報共有を行うため、ディレクトリに基づき、内容によって分類されている文書と、管理が行き届かず正しく分類されていない文書が混在している

先に述べたモデルの妥当性を検証するため、企業内に蓄積された文書の格納形態を調査した。図1に、企業内で製品Aに関する情報が記載された文書の格納形態の例を示す。調査の結果、図1に示すように、製品Aに関する情報が記載された文書は、ディレクトリに基づき、内容によって分類されている場合が多いことが判明した。一方、検索ターム「製品A」を含む文書であっても、図1で示したディレクトリ構造に含まれない文書は、製品Aを中心にしたトピックでない場合が多いことが

判明した。上記の調査結果より、本研究では、先に述べたモデル化は、妥当性があると判断した。

以下、本モデルを踏まえ、同一トピックに関する情報が記載された文書の集合を抽出する手法の考え方を示す。本研究では、同一ディレクトリには内容に関連がある文書が格納されるとして、これらを同一トピックの情報とみなすことができると考えた。図2に、ディレクトリ構造に基づき、同一トピックの文書集合を抽出した例を示す。図2は、製品Aと製品Bに関する情報が記載された文書が、それぞれ別々のディレクトリに格納されていることを示している。

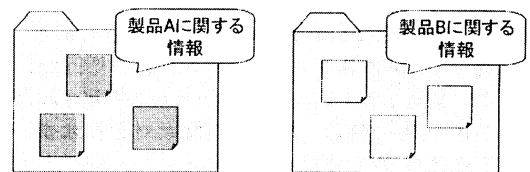


図2 ディレクトリによる抽出の例

本研究では、この考え方に基づき、検索条件に関連の深いディレクトリに含まれる文書は、検索者にとって有用な情報が記載された文書であると評価するランキング方式を提案する。

4.2. 実現方法

本節では、個々の文書の評価に加え、同一トピックに関する情報が記載された文書の集合をディレクトリ構造に着目して抽出し、抽出された文書の集合も評価するランキング方式（以下、文書群評価ランキング方式と呼ぶ）の実現方法を示す。実現方法は、大きく以下に示す4つの処理によって構成される。

(1) ヒット文書取得処理

指定された検索タームを含む文書をヒット文書として文書データベースから取得する。

(2) 文書群別スコア算出処理

ディレクトリに含まれる文書を文書群として抽出し、抽出された文書群に含まれるヒット文書の数を文書群別スコアとして付与する。

(3) 文書別スコア算出処理

文書の更新日時や、文書が格納されているディレクトリの階層の深さを評価して、文書別スコアを算出する。

(4) 検索結果出力処理

上記(2)で算出された文書群別スコアと、上記(3)で算出された文書別スコアに基づき、トータルスコアを算出する。このトータルスコアの降順にヒット文書を並び替え、検索結果として出力する。

5. 評価実験

前述の通り、本研究では、一般的な企業内文書検索における検索対象として、Webサーバと共有ファイルサーバ内の文書を想定している。

本研究で提案する文書群評価ランキング方式は、両サーバに共通するディレクトリ構造に着目したものであるため、両サーバ上の文書に対して、ランキングが可能である。そのため、本方式の有効性の検証にあたり、両サーバを対象とした精度評価が必要であるが、本論文では、まず、Webサーバ上の文書を検索対象とした評価結果を報告する。以下、実施した精度評価の前提条件を説明する。

5.1. 評価環境

自社内で公開されているWebサーバから、製品情報を提供しているサイトを含めた約9万件を対象にして、本精度評価を行った。

5.2. 評価指標

(1) 目的文書の表示順位

3.1節で説明した試用調査と同様にして、「製品情報の調査」を利用ケースとして想定し、各々の製品紹介のトップページを目的文書、製品名を検索タームに設定した。この設定の下、文書群評価ランキング方式とTFランキング方式で得られた目的文書の表示順位をそれぞれ比較した。

(2) 目標表示順位

本精度評価では、検索結果における一画面分の表示件数を考慮して、想定した目的文書の表示順位の目標値を5位以内とした。目標値を5位以内に設定した理由は、検索結果画面のスクロールや次の検索結果画面への移動などの画面操作は、ヒット文書から目的文書を探す上で検索者の負担になるため、目的文書は少なくとも検索結果における一画面分の表示件数以内に含まれるべきであると考えたからである。

5.3. 評価結果と考察

表1に、5.2節の(1)で示した想定条件の下、各ランキング方式で得られた目的文書の表示順位をそれぞれ示す。本結果より、想定した条件の下では、検索結果の上位5位以内に目的文書を表示するという目標を達成することができた。また、TFランキング方式と比べて、目的文書の表示順位を大幅に向上させることができた。以下、本精度評価で得られた結果について考察する。

文書群評価ランキング方式は、検索タームに関する情報を多く含むディレクトリを探した上で、そのディレクトリを代表する文書を探し、目的文書とみなすものである。精度評価の結果より、製品情報の検索においては、#2のサンプルの場合を除いて、本方式は、目的文書の表示順位の向上に効果があったと言える。この効果を検証するため、製品情報が記載された文書が含まれるディレクトリの調査を実施した。

表1 各ランキング方式に対する想定目的文書の順位

#	検索目的	検索ターム	想定した目的文書の順位	
			文書群評価 ランキング方式	TF ランキング 方式
1	製品情報の調査	製品A	1	65
2		製品B	3	505
3		製品C	1	274
4		製品D	1	104
5		製品E	1	541
6		製品F	1	1000 位以下
7		製品G	1	174
8		製品H	1	12
9		製品I	1	52
10		製品J	1	13
11		製品K	1	1000 位以下

調査の結果、製品情報が記載された文書の多くは、個々の製品ごとにディレクトリによって分類されて格納される傾向があることが判明した。そのため、同一ディレクトリに含まれるヒット文書の数の評価することによって、目的文書が含まれるディレクトリを絞り込む効果があったものと考えられる。

表1の#2のサンプルにおいては、同一ディレクトリに含まれるヒット文書の数の評価では、目的文書が含まれるディレクトリを絞り込むことができなかった。本件について、ヒット文書が最も多く含まれるディレクトリを調査した結果、このディレクトリには、製品Bに関連する製品群の拡販に関する文書が大量に格納されていることが判明した。そのため、このディレクトリには、多くのヒット文書が含まれ、高く評価される結果となったと考えられる。

6. まとめと今後の課題

本研究では、企業内においては、検索者に必要とされる文書の多くが、人手によってあらかじめ分類されたディレクトリに格納されているという傾向に着目し、ディレクトリの解析結果から文書の内容を評価する指標を提案した。具体的には、各ディレクトリに含まれるヒット文書の数を書群別スコアとして算出し、ヒット文書の評価に加えるものである。提案した指標を用いた文書群評価ランキング方式では、想定した条件の下では、検索結果の上位5位以内に目的文書を表示するという目標を達成することができた。また、TF ランキング

方式と比べて、目的文書の表示順位を大幅に向上させたことから、文書群評価ランキング方式の有効性を確認することができた。以下、本研究における今後の課題を示す。

(1) 格納形態のバリエーションの検討

精度評価の結果より、表1の#2に示す事例のように、製品情報に関する文書が、個々の製品ごとにディレクトリによって分類されていない場合があることが分かった。今後は、多種類の検索タームで評価実験を行い、本研究で想定外の文書の格納形態について調査を深め、対策を考えていく必要がある。

(2) 更なる精度向上に向けて

提案した指標を用いた文書群評価ランキング方式では、文書の評価とディレクトリの評価の大きく二つからなる。本研究では、ディレクトリの評価に重点をおいたが、今後は、文書の評価に対する検討を深めていく。

文 献

- [1] M. F. Fontoura, A. Neumann, S. Rajagopalan, E. Shekita, and J. Zien, "High Performance Index Build Algorithms for Intranet Search Engines", 30th International Conference on Very Large Data Bases (VLDB'2004), Toronto, Canada, 2004.

- [2] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," In Proceedings of The 7 th International World Wide Web Conference, pp.107-117,1998
- [3] L. Page, "The PageRank Citation Ranking:Bringing Order to the Web,"
<http://dbpubs.stanford.edu:8090/pub/SHOWDoc.Fulltext?lang=en&doc=1999-66&format=pdf&compression=&name=1999-66.pdf>,1998
- [4] 佐々木俊尚 他, "エンタープライズ検索テクノロジー," Computerworld, Jul.2005
- [5] "Google Search Appliance"
<http://www.google.com/enterprise/pdf/datasheet.pdf>