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

A Ranking Method for Enterprise Documents Search Based on Directory Structures

四ッ谷 雅輝† 松林 忠孝‡ 弥生 隆明‡
野田 十悟‡ 吉田 豊‡

Masaki YOTSUTANI Tadataka MATSUBAYASHI
Takaaki YAYOI Jugo NODA Yutaka YOSHIDA

企業内に蓄積された膨大な数の文書から，必要な情報を取得する方法として，検索システムを利用する方法がある．企業内における検索システムの利用ケースでは，漠然としたキーワードが少ない個数しか指定されないようなケースも考えられる．このような場合，検索者にとって必要な情報を含んだ文書を検索結果の上位に表示できず，必要な情報を取得しにくいことがある．そこで，本研究では，企業内で必要とされる文書の多くは，人手によって予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，ディレクトリに含まれるヒット文書の数を取得して，文書の重要度を算出する上での一指標に用いることを提案する．本手法は，ディレクトリ構造に着目するものであるため，Web サーバ上の文書に限らず，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．

We can use a search system when we want to obtain useful information among a large amount of documents in an enterprise. We can think that users often submit few ambiguous keywords to the enterprise search system. In such case, users cannot obtain useful information because the documents that contain useful information for users are not ranked highly in the search results. In this paper, we propose a measure for computing the importance of documents, focusing on the fact that a lot of documents utilized in an enterprise are stored in the directories classified manually. This measure is computed based on the number of hit documents in the directory. We can apply our proposed method to searching documents not only on Web servers but also on shared file servers.

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．検索者の要望を満たす検索システムを実現するためには，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．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）の一覧を取得するステップ，(3)取得したヒット文書の一覧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な検索を実現するためのインデクスを作成する技術の開発などを行ってきた．

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．

特に，企業内における検索者は，期日が定められた業務に必要な情報を探すために，検索システムを利用するという性質上，検索に時間や手間をかけることが難しい．そのため，指定される検索ターム数が少ないといったような簡単な検索条件の場合であっても，検索者にとって必要な文書を検索結果の上位に表示できるランキング技術が求められると考えている．通常，漠然としたキーワードが少ない個数しか指定されない場合，大量の文書がヒットし，検索者にとって必要な情報を含んだ文書を検索結果の上位に表示できず，必要な情報を取得しにくいことがある．そこで，本研究では，企業内で必要とされる文書の多くは，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，ディレクトリに含まれるヒット文書の数を取得して，文書を評価する上での一指標に用いることを提案する．

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，企業内文書の特徴に着目し，文書を評価する上での一指標を提案する．４章では，提案に対するアプローチおよびその実現方法について説明する．５章では，提案方式に対する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，Google§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重要度を算出する方式である．算出された重要度は，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Webページ

† 正会員 （株）日立製作所 ソフトウェア事業部 Hitachi, Ltd. Software Division m-yotsutani@itg.hitachi.co.jp
‡ 非会員 （株）日立製作所 ソフトウェア事業部 Hitachi, Ltd. Software Division {tadamats,t-yayoi,noda_jy, yosida_y}@itg.hitachi.co.jp

§ Googleは，Google Inc. の登録商標です．

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，目的文書に高い重要度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文書管理システムなどに格納された文書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（以下，TF 値と呼ぶ）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. 本研究の着目点

企業は，多数の部門から構成されており，各部門には部門特有の情報が蓄積されていることが多い．また，部門内に蓄積された情報を有効に活用するため，文書をディレクトリによって分類して管理していることが考えられる．例えば，製品情報は，製品名ごとにディレクトリによって分類され，プロジェクトの関係書類は，プロジェクト名ごとにディレクトリよって分類されることが考えられる．

しかしながら，このように，ディレクトリによって文書が分類されていたとしても，これまでの検索システムでは，文書の格納形態までは考慮されておらず，分類された文書と未分類の文書を区別して評価することは難しい．そこで，本研究では，企業内で必要とされる文書の多くは，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，ディレクトリに含まれるヒット文書の数を取得して，文書を評価する上での一指標に用いることを提案する．次章では，そのアプローチと実現方法について説明する．

4. 方式検討

4.1 アプローチ

本研究では，検索者にとって有用なディレクトリを，検索タームに関連した情報を多く含むディレクトリと考えた．そして，検索タームに関連した情報を多く含むディレクトリを，ヒット文書が多く含まれるディレクトリとみなすアプローチを検討した．図 1 に，本アプローチの概念図を示す．

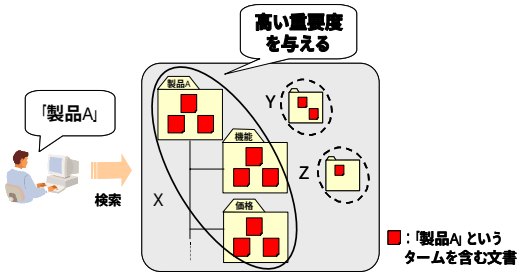

図 1　本アプローチの概念図

Fig.1 Concept chart of our approach

図 1 では，検索タームとして「製品Ａ」が指定された場合，ディレクトリＹやディレクトリＺのように，含まれるヒット文書が少ないディレクトリよりも，ディレクトリＸのようにヒット文書を多く含むディレクトリに格納されたヒット文書の方が，検索者にとって有用であるとみなし，高い重要度を与える考え方を示すものである．

4.2 実現方法

本節では，個々の文書の評価に加え，ディレクトリの評価も加味したランキング方式（以下，提案方式と呼ぶ）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) ディレクトリ別スコア算出処理
　ヒット文書を多く含むディレクトリほど高いスコアを付与する．
(3) 文書別スコア算出処理
　検索タームを多く含む文書ほど高いスコアを付与する．
(4) 検索結果出力処理
　上記(2)で算出されたディレクトリ別スコアの降順にディレクトリを選択し，選択されたディレクトリの中から上記(3)で算出された文書別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．

上記(1)～(4)のうち，提案方式の中核を担う(3)「ディレクトリ別スコア算出処理」について，処理ステップを以下に示す．

[Step1] ヒット文書の URL（あるいは，ファイルパスなど）を取得する
[Step2] 取得したヒット文書の URL を解析し，ヒット文書が含まれるディレクトリを判別する
[Step3] 取得したヒット文書の URL に基づき，各ディレクトリに含まれるヒット文書を判別する
[Step4] 各ディレクトリに含まれるヒット文書の数を計数する

なお，上記[Step4]に関しては，計数の対象とするヒット文書を，設定した TF 値の閾値を超えるものに限定することで，ヒット文書を厳選し，ノイズを除外してディレクトリの評価を行うことも可能である．

図 2 に，TF 値が「5」，「4」，「1」の 3 つのヒット文書が含まれるディレクトリに対して，TF 値の閾値「3」を設定した場合のディレクトリ別スコアの算出例を示す．図 2 では，TF 値の閾値「3」を超える，TF 値が「5」，「4」のヒット文書のみが計数の対象となり，ディレクトリ別スコアとして，「2」が算出されることを示している．

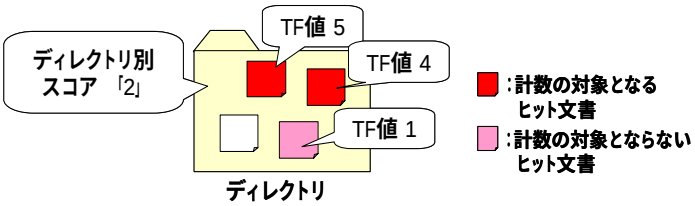

図 2　閾値を TF 値 「3」に設定した場合の算出例

Fig.2 Example of a computing method when threshold is set to TF value "3"

5. 評価実験

本研究における提案方式は，ディレクトリ構造に着目したものであるため，Web サーバに限らず，共有ファイルサーバ上の文書に対しても，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．

5.1 評価指標

評価指標としては，設定したキーワードに対する目的文書の表示順位を用いるものとし，提案方式と TF ランキング方式によって得られた結果をそれぞれ比較した．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．なお，本精度評価には，自社内で公開されている Web サーバ約 9 万件を用いた．

5.2 評価方法

本精度評価では，企業内文書検索の利用ケースを想定するため，以前より運用されていた自社内の Web サーバを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．これにより，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．

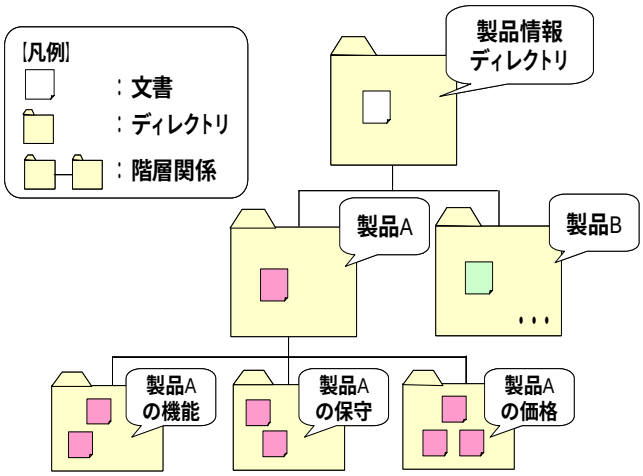

図 3　製品Ａに関する文書の格納形態の例

Fig.3 Example of a storage form about documents of concerning product A

また，自社内の Web サーバにおける製品情報が提供されているサイトを調査した結果，図 3 に示すように，各部門で蓄積される文書と同様にして，ディレクトリによって文書が分類される場合が多いことが分かった．そのため，ディレクトリの評価を加味できる提案方式は，適用による効果が高いと判断し，精度評価を実施した．以下，本精度評価で設定した目的文書と検索タームをそれぞれ説明する．

(1) 目的文書

「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，本評価では，製品紹介ホームページのトップページを，製品に関して様々な情報が記載されたページであるとして，目的文書に想定した．

(2) 検索ターム

基本的に，検索システムを利用する場合，検索者は探している情報について，詳細な内容を把握できているケースは少ないと考えられる．そのため，分からない情報を得るために検索システムを利用するという状況の中では、漠然としたキーワードが少ない個数しか入力されない場合が多い．そこで，本研究では，検索タームとして製品名の一語のみが入力された場合を想定して評価した．

5.3 評価結果と考察

表 1 に，5.2 節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．なお，表中の（）内は想定した目的文書を含むディレクトリの順位（以下，ディレクトリ順位と呼ぶ）を示す．

表 1　各ランキング方式に対する想定目的文書の順位

Table1 The order of the assumption purpose document by each ranking method

#	検索ターム	ヒット文書数	想定した目的文書の順位	
			TF ランキング方式	提案方式
1	製品A	2,028	65	2 (1)
2	製品B	2,925	274	10 (3)
3	製品C	870	104	68 (4)
4	製品D	4,269	505	307 (2)
5	製品E	2,288	541	1 (1)
6	製品F	589	174	5 (1)

本結果より，想定した条件の下では，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．

(1) 表示順位向上の理由について

比較した TF ランキング方式による検索結果では，製品広告メールの送信履歴や製品発注フォーマットというような，検索タームが多く含まれる文書が検索結果の上位に表示されていた．これらの文書が含まれるディレクトリを調査した結果，製品仕様などに従って，ディレクトリが階層関係を形成していることが判明した．この中で，TF ランキング方式によって上位に表示されていたヒット文書は，下位階層のディレクトリに含まれる場合が多く，その結果，ディレクトリに含まれるヒット文書数は少なくなり，検索結果の表示順位を低く抑えることができたと考えられる．

(2) 上位ディレクトリの内訳について

検索結果の上位に表示された文書が格納されたディレクトリの内訳は，大きく以下の３つに分けることができる．

(a) 製品ディレクトリ
(b) 旧バージョンの製品ディレクトリ
(c) 製品広告メールのバックナンバーの保管用ディレクトリ

上記(a)に関しては，製品情報を調査している検索者にとって，有用なディレクトリであるのは明らかと考える．また，(b)(c)に関しても，上記(a)と比較すると優先順位

は劣るものの，いずれのディレクトリも製品情報に関連しており，検索者にとって有用とみなすことができる．このことより，提案方式によって，検索対象に含まれる文書のうち，分類された文書と，未分類の文書を識別できる見込みを得た．

一方，上記(a)〜(c)を比較して分かるように，同じ「製品情報」のディレクトリであってもバージョンの違いを識別することや,「製品情報」のディレクトリと「広告メール」のディレクトリといった，ディレクトリによる分類の目的の差異を識別することは困難であることが判明した．

以上のような，ディレクトリによる分類の目的の差異を識別するための対策案としては，文書の更新日時といったディレクトリに含まれる文書の特性も評価の観点に含め，多角的な評価を行うことが考えられる．

(3) 表示順位が低い場合の理由について

4.2 節で述べた通り，現段階では，ディレクトリ順位の降順にディレクトリを選択し，最上位階層のディレクトリに含まれるヒット文書を TF 値の降順に表示するものである．そのため，最上位階層のディレクトリに多数のヒット文書が含まれる場合，検索結果の上位には，本ディレクトリに含まれるヒット文書しか表示されないという現象が起きてしまうことが判明した．これに対する対策案としては，検索結果に表示されるヒット文書が，特定のディレクトリに含まれるものにならないように，予め 1 ディレクトリあたりに表示するヒット文書数を設定しておくことが考えられる．

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，人手によって予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，ディレクトリに含まれるヒット文書の数を取得して，文書の重要度を算出する上での一指標に用いることを提案した．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ディレクトリ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. 提案方式は，想定した条件の下では，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，そ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．

本研究では，ディレクトリを評価する際にヒット文書の数を用いるものとしたが，ディレクトリに含まれる文書に対するヒット文書の割合などを考慮することも考えられる. また，ディレクトリを評価する際に用いるヒット文書の厳選方法に関しても，本研究で示した検索ターム数以外の観点からも検討する必要がある．

これに加え，ディレクトリの評価に関して，ヒット文書以外の観点も考慮に入れ，ディレクトリの階層構造の深さやディレクトリに含まれるサブディレクトリの数など，多角的な評価について検討していく．

[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

四ッ谷 雅輝 Masaki Yotsutani
2003 年 北海道大学大学院工学研究科電子情報工学専攻修了. 同年,（株）日立製作所入社．入社以来，検索システムの精度向上に関する研究に従事．文書要約や文書分類など、自然言語処理を応用した技術にも興味を持つ. 情報処理学会正会員. 日本データベース学会正会員.

松林 忠孝 Tadataka Matsubayashi
1996 年 東京工業大学総合理工学研究科電子システム専攻修了．現在,（株）日立製作所ソフトウェア事業部に勤務．文書検索システムの研究開発に従事．情報処理学会正会員.

弥生 隆明 Takaaki Yayoi
1998 年 筑波大学大学院　理工学研究科修了．現在,（株）日立製作所ソフトウェア事業部に勤務．文書検索システムの研究開発に従事．情報処理学会正会員.

野田 十悟 Jugo Noda
1999年大阪教育大学大学院教育学研究科総合基礎科学専攻修了．同年,（株）日立製作所入社．入社以来，検索エンジンの開発に従事.

吉田 豊 Yutaka Yoshida
2000 年 名城大学大学院理工学研究科電気電子工学専攻修了. 現在,（株）日立製作所ソフトウェア事業部に勤務．文書検索システムの研究開発に従事．情報処理学会正会員.

** Googleは, Google Inc. の登録商標です.