2008－DD－68 (5)
2008／9／26

文書の編集関係を利用した文書の探索方法

長谷川雅一† 綱川 光明† 鷲崎 誠司†
NTT サイバースペース研究所†

あらまし 近年，企業活動の中で様々な文書が作成され爆発的に増え続けている．有用な情報を迅速に入手し，それらを参照，あるいは流用して適切な文書を作成することが企業の生命線であり，これら有用な情報は既存文書群に遍在していることが多い．Web 上の情報検索技術は Google を筆頭に様々なアプローチで研究開発が進められている一方で，社内文書からの情報探索については，Web 上の検索技術を流用するだけでは，検索結果の適合率が低くなることが多い．本提案では，社内の共有文書群の中から，文書間の派生関係を抽出し，この関係性を用いて適切な文書を探索する方式を提案する．
キーワード 文書検索，文書探索，派生関係，企業内文書

Document exploring method using change tracking

Masaichi Hasegawa†, Mitsuaki Tsunakawa†, Seiji Susaki†
NTT Cyber Space Laboratories†

Abstract In recent years, the number of various documents increases explosively as a result of company's activities. It is vital for workers in enterprises to build a proper document promptly by reusing useful information stored in existing documents. Various techniques of information retrieval for the web documents have been conducted by Google and others, however if those techniques are applied to the documents in enterprises, the precision rate of the search result is not sufficient. In this paper, we describe an information retrieval technique for enterprise documents; It improves the precision rate by taking the extracted document histories into account.
Key words Document retrieval, Document search, Derivation relation, Document in enterprise

1. はじめに

近年，IT 化やストレージ価格の低廉化などを受けて，あらゆる文書が目的や用途別に作成され，企業内に蓄積され続けている．これら文書は，文書管理規定が及ばない草案やメモなどの個人管理文書の範囲内においては，各人に整理が委ねられ，統制されずに文書が増え続けているのが現実である．

企業活動においては，これら文書の中から，有用な情報を迅速に入手，活用して適切な文書を作成し続けることが生命線となっている．これに対し，Web 上の情報検索技術を利用するのが一般的となっている．しかし，当該技術は文書を単体とし，空間軸のみを対象としている．本来，それだけでは不十分であり，文書の時間軸方向の変遷や，異なる文書間の繋がりを辿れるようにし，文書への到達率を高めることが重要であると考える．

従来技術として，Namazu[1]などの全文検索技術がある．これらは，形態素解析や N-gram の方式を用いてインデキシングし，TF-IDF[2]などのランキング技術を用いてヒットした文書をソートし，返却する技術である．しかし，インデキシング時点の文書集合のみを検索対象とし，文書の更新履歴を辿ったり，文書間の関係性を辿ることはできない．

Microsoft の SharePoint[3]や，Subversion[4]など，オンライン型の文書のバージョン管理システムが提案されている．これは，ユーザの操作(チェックイン/チェックアウトなど)により，文書の改版履歴を検知し，検索可能とするものである．柔軟な検索を実現するとともに，文書の更新履歴を辿ることは可能であるが，ユーザに高度な操作手順を強いること，オフラインでの文書間の流用履歴は管理していないので文書間の関係性を辿れないこと，などの問題があった．

NEC のタイムトラベル検索技術[3]が提案されている．これは，文書の更新差分を管理し，所望

の時点の共有文書ファイルサーバの状態を復元し，それに対して検索する技術である．しかし，異なる文書間の繋がりを管理していないため，文書の更新履歴は辿れるが，文書間の繋がりを辿ることはできない．

本稿では，文書のライフサイクルに着目し，文書の更新履歴や，文書間の情報の流用履歴をモデル化し，これらの情報を活用して文書を検索するとともに，文書間をブラウジングする(辿る)手法により，目的の文書への到達度を高め，効率的に探索する手法を提案する．

2. 前提条件

2.1. 検討の前提

今回の検討を進めるにあたり，以下を前提条件として設定している．

- 文書構造をモデル化するにあたり，ファイルフォーマットに特化させず，企業内の一般的な文書に適用できること．
- チェックイン/チェックアウトなどのユーザ操作を必要としないこと．
- 既存環境への負担の少ない導入を考慮し，ファイルサーバやユーザの文書生成環境への追加ソフトの導入は不要とすること．

2.2. 文書のライフサイクル

文書のライフサイクルを整理し，モデル化した結果を図1に示す．文書（ファイル）に対する操作は，以下の5つに分類できる．

(1) 作成(新規作成)
(2) 作成(ファイルコピー)
(3) 削除
(4) 更新(上書き保存)
(5) 更新(別名保存)

ファイル名が変更されない操作を改版と呼ぶこととする．上記では，(1)のあと，(4)を繰り返し，最後に(3)とする操作が文書の改版である．なお，(4)や(3)の操作はなくても良い．

文書を作成や更新するにあたり，他の文書の全部または一部を流用することを派生と呼ぶこととする．上記では，(2)，(5)の操作が文書の派生にあたる．

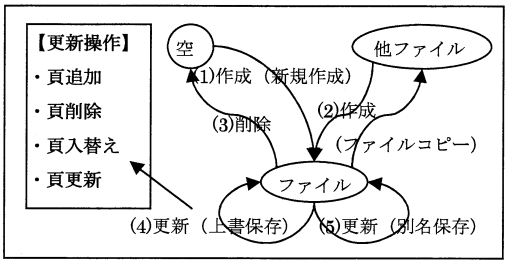

図1. 文書のライフサイクル・モデル

2.3. 文書の改版判定

一般的に文書の内容を変更すると，ファイルのタイムスタンプとファイルサイズが変わることに着目して改版を判定する．

2.4. 文書の派生判定

(1) 文書構造のモデル化

流用されることが多いPowerPointを例に文書構造をモデル化する．

- 文書は，1個以上の頁から構成される．
- 1頁には，0or1個のタイトルと，0個以上の頁テキストから構成される．
- 頁は，順序性を保持するものとする．
- 頁構造は，図2の様に大きく頁タイトルエリアと頁テキストエリアに分割できる．
- 先頭頁の頁タイトルは，文書タイトルとする．

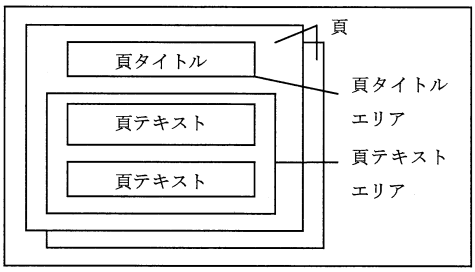

図2. 文書構造と頁構造

頁タイトルとは，頁の内容を表すタイトル・コンテンツであり，頁タイトルエリアに入力されている．

頁テキストとは，頁の内容を具体的に記述しているコンテンツであり，頁テキストエリアに入力されている．

なお，Microsoft の Word であれば頁は章に，同 Excel であれば頁はワークシートに対応させ

て，同様のモデルが適用可能と考える．

(2) 文書の派生判定

文書の派生は，元文書の何らかの痕跡を残すものと考えられる．ファイルの更新操作と文書構造との関係を整理したものが表 1 である．なお，文書に対する操作 (2)作成(ファイルコピー)は，ファイル更新の前作業と考え，今回は表 1 の整理からは除外した．

表 1. 派生文書派生時の更新操作種別と他情報の関係

更新操作＼文書構造		文書タイトル	頁タイトル	頁テキスト
頁追加		不変	増加	不変
頁削除		不変	減少	不変
頁入替え		不変	不変	不変
頁更新	頁テキスト	不変	不変	増減
	頁タイトル	不変	変化	不変
	文書タイトル	変化	不変	不変

一般に，文書編集活動において，元文書の痕跡を全て消去されることは考えづらい．表 1 中の「増加」，「減少」，「変化」の量を収集し，閾値として判定することでファイル派生を推定する．

3. 提案手法

3.1. 時系列アーカイブ方法

文書空間状態の状態を再現するために共有文書空間の文書を定期的に収集する．

文書のライフサイクルを利用するには，文書に対する操作を取り出す必要がある．前回の収集契機との違いをチェックし，文書の作成，更新，削除を判断する．また，最近の共有ファイルサーバのファイルシステムでは，文書を作成，更新，削除した操作を通知する仕組みがあるものもある．この場合は，当該仕組みを利用して，定期収集と定期収集の間の文書共有空間の状態変更を文書に対する操作と考えて文書に対する作成・更新・削除操作契機で文書を収集する．

収集した文書を更新日時の古い順に並び替えて，最も古い文書から逐次，以下の a)〜c)の処理を文書がなくなるまで繰り返す．

以下，文書の「ファイル名」，「タイムスタンプ」(「作成日時」，「更新日時」)，「ファイルサイズ」と「ファイルパス」を外部属性情報と呼ぶこととする．「ファイルパス」は，文書が文書ファイルサーバに保存されているパスである．また，文書の「構造」と「文字列」を内部属性情報と呼ぶこととする．

a) 文書をアーカイブする

1) 文書識別子は，「ファイルパス」と「ファイル名」を組み合わせて作成し，「文書識別子」から文書を取得できる様に文書の実体を蓄積する．

2) 改版判定のために「文書識別子」から文書の外部属性情報を取得できる様に「文書識別子」と文書の外部属性情報を関連付けして蓄積する．

3) 派生判定のために「文書識別子」から文書の内部属性情報を取得できる様に「文書識別子」と文書の内部属性情報を関連付けして蓄積する．

4) 文書からテキスト（単語）を抽出し転置インデックスを作成する．

b) 改版を判定する

同一文書識別子，かつ，収集文書の「タイムスタンプ」と「ファイルサイズ」が前回収集時と異なる場合は，文書が改版されたと判断し，「文書識別子」と「タイムスタンプ」を改版履歴に蓄積する．

c). 派生を判定する

表 1 の内部属性情報を用いて派生判定するが，文書タイプ（説明用文書，帳票）によって判定結果が異なる可能性があるので，内部属性を組み合わせて 4 種類の派生判定ルールを準備し，夫々，独立に判定する様に実装した．

派生判定ルール 1) 同一頁の有無で判定
（表 1 の点線）

全ての内部属性を用いて判定する．

判定方法：同一頁タイトル，かつ，同一頁テキストを持つ頁が一つ以上ある場合は，派生関係ありとする．

想定ノイズ：コピー元の文書までは特定できない．

派生判定ルール 2) 類似する頁数で判定

（表 1 の一点破線）

頁テキストを用いて判定する．

判定方法：同一頁テキストが一定数（y）以上ある頁を類似頁とみなし，類似頁が一定数（x）以上ある場合は，派生関係ありとする．

想定ノイズ：頁テキストが少ないと誤判断する．

派生判定ルール 3）同一頁タイトル数で判定

（表 1 の破線）

頁タイトルを用いて判定する．

判定方法：同一頁タイトルが一定数（x）以上ある場合は，派生関係ありとする．

想定ノイズ：会議など統一ストーリ文書は派生と誤判断される．

派生判定ルール 4）同一文書タイトルで判定

（表 1 の実線）

文書タイトルを用いて判定する．

判定方法：同一文書タイトルがある場合は，派生関係ありとする．

想定ノイズ：統一形式の帳票類は派生と誤判断する．

3.2. 検索方法

文書を検索する時は，目的の文書の「キーワード」とファイルの存在期間を「検索期間」として指定する．内部処理では，指定された「検索期間」に存在した文書に限定し，「キーワード」を用いて全文検索を実行し，ヒットした文書識別子と当該文書の存在期間のリストを取得する．文書識別子から，その改版履歴（「タイムスタンプ」）を取得するインタフェースを準備した．

また，派生関係を辿るため，文書識別子が指定された場合に，その派生元・派生先文書識別子を取得するインタフェースを準備した．

3.3. 検索結果表示方法

検索結果は，2 ステップで表示する．

(1) 検索にヒットした文書識別子をリスト表示する．

(2) 文書識別子が指定されると，その改版履歴，派生元文書の文書識別子，派生先文書の文書識別子を取得し，タイムライングラフで表示する．タイムラインの表示例を図 3 に示す．

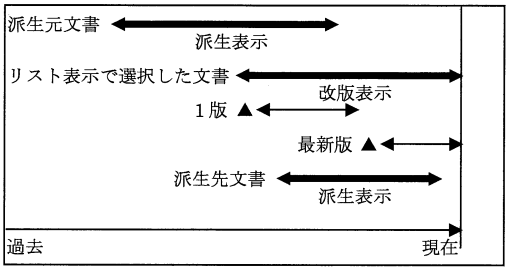

図 3. 検索結果のタイムライン表示例

さらに，タイムライン表示から派生元文書や派生先文書を選択すると，その文書を中心として，タイムライン表示を遷移することとし，派生関係を辿れるようにした．

4. 評価

派生判定ルールの有用性評価のため，社内の技術検討目的に実際に作成された文書に対して派生判定ルールを適用し評価を実施した．

(1) データセット

PowerPoint ファイル 18 個（文書タイプ：説明資料及びその検討資料）で，正解を図 4.に示す．正解数は，派生関係数が 22 件，派生関係の重み付けは行わない．

凡例 A1，A6，A10，A16，A17，A18 は，説明資料であり，これ以外の凡例は，検討資料である．ファイルのタイムスタンプは，A1 が最も古く，次に古いのは，A2 となり，A18 が最も新しい．なお，図 4 中，グラフの高さには意味はない．視認性を高めるために調整した．

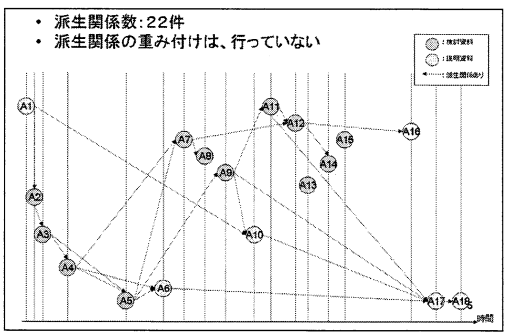

図 4. 文書の派生関係の正解

(2) 測定パターン

測定パターンを表 2 に示す．派生判定ルールのパラメータである，派生判定ルール 2 の x，y と派生判定ルール 3 の x を変化させ，適合率・再現率を測定した．なお，適合率と再現率の式は以下の通りである．

適合率＝抽出した正解数／抽出した派生数

再現率＝抽出した正解数／全体正解数

表 2. 測定パターンとパラメータ

#	測定パターン	
1	派生判定ルール 1	同一スライドあり
2	派生判定ルール 2	テキスト 25%一致，頁 25%一致
3		テキスト 50%一致，頁 25%一致
4		テキスト 75%一致，頁 25%一致
5		テキスト 25%一致，頁 50%一致
6		テキスト 50%一致，頁 50%一致
7		テキスト 75%一致，頁 50%一致
8		テキスト 25%一致，頁 75%一致
9		テキスト 50%一致，頁 75%一致
10		テキスト 75%一致，頁 75%一致
11	派生判定ルール 3	頁タイトル 25%一致
12		頁タイトル 50%一致
13		頁タイトル 75%一致
14	派生判定ルール 4	同一文書タイトルあり

(3) 評価結果

測定結果を表 3 に示す．

派生判定ルール 2（#10）は適合率が，派生判定ルール 2(#2)は再現率が高いため，利用用途により使い分けが必要である．

派生判定ルール 2，3 での正解抽出に関して包含関係ある．(正解[#3-#10]⊂正解[#2]，正解[#12, #13] ⊂正解[#11]（但し，不正解も抽出される))

派生判定ルール 2 と 3 は，適合率・再現率のバランスしている#5 と#12 を代表点としてプロットしたのが図 6 である．

表 3. 評価結果

#	適合率	再現率	全体正解数	抽出した派生数	抽出した正解数
1	0.348	0.727	22	46	16
2	0.309	0.773	22	55	17
3	0.317	0.591	22	41	13
4	0.481	0.591	22	27	13
5	0.565	0.591	22	23	13
6	0.571	0.545	22	21	12
7	0.667	0.455	22	15	10
8	0.533	0.364	22	15	8
9	0.571	0.364	22	14	8
10	0.800	0.364	22	10	8
11	0.500	0.545	22	24	12
12	0.625	0.455	22	16	10
13	0.750	0.273	22	8	6
14	0.625	0.227	22	8	5

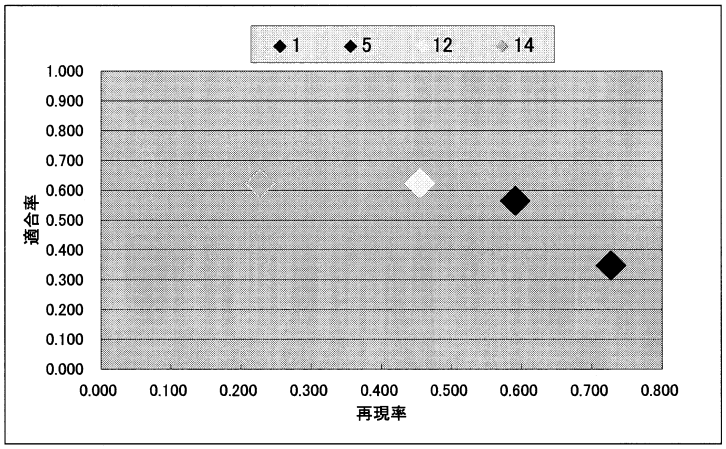

図 6 適合率・再現率

5. さいごに

本稿では，文書の更新履歴や情報の流用履歴を利用する文書検索・探索方法について提案した．

本方法により，文書の過去の版に遡って検索できるとともに，派生関係を辿ることで所望の文書への到達度を高めることができる．

今後は，以下の 2 点について研究していきたい．

- 適合率を重視すると「派生関係あり」となるはずの文書が抽出されないため，文書の関連性を見出せない．これについては，Vector モデルの適用や機械学習による類似判定モデル（SVM, LSI, etc)を併用する方法が考えられる．
- 今回は社内の技術検討目的の文書を用いて評価したが，他の性格の異なる文書タイプ(意志決定会議用，帳票，議事録，など)を用いて派生判定ルールの検証が必要である．

参考文献

[1] Namazu Project：
http://www.namazu.org/index.html.ja

[2] 杉山 一成，波多野 賢治，吉川 正俊，植村 俊亮：ハイパリンクで結ばれた隣接ページの内容に基づく Web ページのための TF-IDF 法の改良(情報検索・文書分類)，電子情報通信学会論文誌，Vol.J87-D-I, No.2(20040201) pp. 113-125

[3] Microsoft®：SharePoint
http://office.microsoft.com/ja-jp/sharepoint server/default.aspx

[4] CollabNet, Inc.：http://subversion.tigris.org

[5] 菅 真樹，鳥居 隆史，梶木 善裕：過去に遡った検索を実現する情報検索システムの提案，電子情報通信学会総合大会論文集 D-4-8 pp32 (2007)