M-129

ドキュメントアクセス履歴からのドキュメント関連抽出手法

A Method for Extracting Document Correlations from Document Access History

駒村　典之†　藤原　彰彦†　加藤　裕之†　城所　和明†

Noriyuki Komamura　Akihiko Fujiwara　Hiroyuki Kato　Kazuaki Kidokoro

1．はじめに

オフィス環境でのドキュメントの電子化が進むことにより、オフィス業務の多くではメールや電子ドキュメントのやり取りが業務の内容のほとんどを占めるようになっている。このような環境では、業務に関連するドキュメントの収集やドキュメントに記載された情報の確認など情報を再利用する為にかかっている時間をいかに減らし、ユーザが本来の業務であるドキュメントの作成や判断に時間を使えるようにできるかが業務の効率化に大きな影響を与える。オフィスドキュメントの再利用を促進する方法として、我々はドキュメントの内容ではなく、そのドキュメントのアクセス履歴に注目し、ドキュメント間の関連を辿っていくことで目的のドキュメントを探し出す方法を検討している。履歴情報には、ユーザが実業務を進める最中に行った判断や、ドキュメントの選択などに関する情報が残されていると考えられる[1]。 履歴情報からユーザの処理手順やドキュメントの因果関係に関する情報を抽出して利用することが出来れば、適切なキーワードの情報が無い場合やドキュメントの内容について多くを知らない場合でも、必要とするドキュメントを効率よく引き出すことが可能となる。

2．履歴解析手法

あるユーザが、製品の仕様書作成業務，市場調査報告書作成業務，研究費管理業務といった複数の「業務」を並行して行った日のドキュメントアクセス履歴には、その複数種類の業務に関する履歴が混在しており、履歴の中では同じ業務であっても時間的に離れた幾つかの小さな「作業」に分かれてしまう。このような履歴情報を利用してドキュメントの関連を抽出する為に、ドキュメントアクセス履歴において作業の変わり目を見つけ履歴を分割し、同じ業務毎に作業をまとめる必要がある。この作業の変わり目について、ドキュメント操作の傾向に関する経験則より次のような仮定をたてた。

仮定 A. 複数の業務を同時並行して行っているユーザーでも、ある短い時間範囲では集中して単独の作業を行っている

仮定 B. 作業の目的が変わる時には、保存や印刷といったような特徴的な操作を行う。

仮定 C. 同じ目的の業務に属する作業の中では似通ったドキュメント群に対するアクセスが発生する。

これらの仮定にしたがってドキュメントアクセス履歴に対して以下の作業を行う。

・クラスタ作成
・ケース作成

† 東芝テック株式会社
デジタルソリューション研究所

クラスタとは上記の例で「作業」にあたるものであり、ドキュメントアクセス履歴を作業の変わり目で分割したものである。この処理によりドキュメントアクセス履歴は複数のクラスタに分割される。作業の変わり目の判定は仮定 B に従い、「保存した」「送信した」などの作業の切り替わり時に発生する特徴的なイベントにより行う。

ケース作成とは、この同じ業務に関係のあるクラスタを「ケース」と呼ぶまとまりに集約する処理であり、一つのケースには同じ業務に属すると考えられる作業が含まれている。前述の仮定 C に従い、クラスタの中身を比較してクラスタの中に同じドキュメントが複数存在する場合や同じ名前のドキュメントを上書きしているような場合では、それらは類似するクラスタであると判断し同じケースにまとめる。

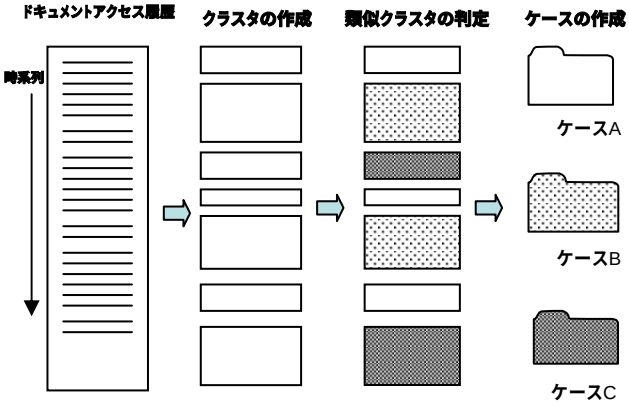

図 1 ケース作成までの流れ

次のステップでは出来上がったケースからドキュメントの関連を抽出する。本手法でこのケースから抽出するドキュメントの関連は、時間的に「ドキュメント A が開かれた」後に「ドキュメント C が保存された」という履歴が残っていた場合、「ドキュメント C の作成の際にドキュメント A を参照した」という参照/派生関係のみを用いる。

今回解析対象としたドキュメントアクセス履歴とは、Microsoft Office、Microsoft Internet Explorer、Lotus Notes のアプリケーションに PlugIn モジュールを追加し、日常の業務を行う最中にそれぞれのアプリケーションでドキュメントに対する「開いた」「保存した」「印刷した」「送信した」などのイベントを取得し、アクセス履歴を時系列で記録したものである。尚 Lotus Notes のドキュメントとはメール文書と掲示板やディスカッション共有ドキュメントなどを含んでいる。収集された履歴は、クラスタ作成、ケース作成、ドキュメント関連抽出の機能を持つ「履歴解析エンジン」で解析される。

図 2 にケースから抽出したドキュメントの関連の一部を図示した。ノードの中にはドキュメントのタイトルとドキ

ュメントの作成者が書かれている。凡例にも示したように、ノードとノードを結ぶ矢印はドキュメントの参照、派生関係を示しており、城所が「修正依頼リスト 030603.doc」を作成する際に、自身の作成した「修正依頼リスト 030520.doc」や駒村の作成した「Open Items excel sheet」を参照したという関係がわかる。矢印の線が太いほど、参照頻度が高くなっていることを示しており、城所は「修正依頼リスト 030603.doc」を作成する際に「Alpha10 英語 UI 仕様書.doc」を頻繁に見ていることが、これによりわかる。この関連を用いることで「修正依頼リスト 030603.doc」からその元になったドキュメントを見つけることが出来る。

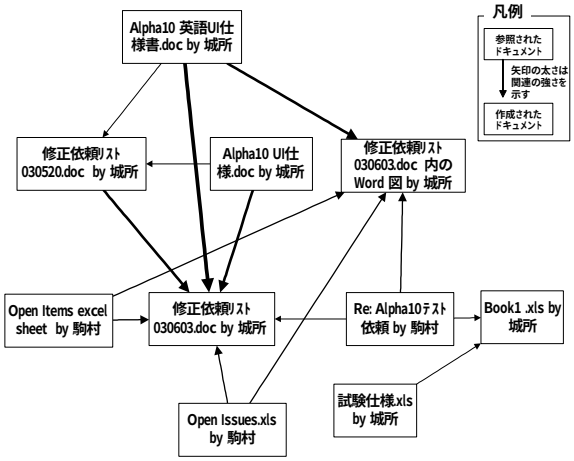

図 2. ドキュメントの関連

3．精度評価

まず、業務履歴から得られる関連情報の特徴を確認する為にユーザ自身の手によってケース(=業務履歴) を作成し、作成した業務履歴から関連を抽出してみた。ドキュメントの関連抽出に関する評価は 2 名の計 6 日分の履歴に対して、実際にドキュメントにアクセスしたユーザが関連するドキュメントと認識している関係のうち本手法でどれ程の関連が抽出されたかを「再現率」として、また抽出された関連にユーザが認識する関連がどの程度含まれているかを「適合率」として数値化した。評価によって得られた、再現率と適合率の平均値を以下に示す。

- 抽出された関連の再現率　　45%
- 抽出された関連の適合率　　80%

全文検索エンジンの再現率が一般に 20～30%で、適合率が 40～50%であることを考慮すると業務履歴抽出ではどちらも高い値が得られる。特に適合率が高く、結果に含まれる関連にノイズが少ないのが特徴である。再現率に注目し、抽出できなかった関連を分析すると、参照/派生以外の関係が多数を占めている。ドキュメントアクセス履歴に見られるドキュメント関係を考察した結果、平均して、同時に参照する頻度の高いドキュメント関係をまとめた「共起関係」が履歴の中に含まれる割合は 1%、同じメールスレッドに属する関係の割合は 20%、Web や共有 DB 等において、既に同じカテゴリに分類された関係の割合は 33%である。共起関係が思いのほか低い割合ではあるが、評価対象の履歴を複数ユーザのドキュメントアクセスを含んだものや、より長い期間の履歴にすることで、解析傾向は変化すると考えられる。現在の手法で抽出される「参照/派生関係」以外のドキュメント関係を関連の抽出に考慮することで、再現率はおよそ 70%程度に向上する。

次に、ドキュメントアクセス履歴からドキュメントの関連抽出までの全ての作業を解析エンジンで行った場合についての評価を行う。「クラスタの作成」「ケースの作成」「ドキュメントの関連抽出」の作業それぞれに対する精度を求めた。精度評価は、作業を行ったときに「正解」となる正しい結果をあらかじめ準備しておき、解析エンジンの結果をそれと比較することにより行った。この「正解」は、ドキュメント操作を行ったユーザ自身がドキュメントアクセス履歴を手作業で解析したものである。これによって得られた各精度の平均値は

- クラスタ作成精度　　50 %
- ケース作成精度　　75 %

である。これらの値から、解析エンジンが正しい関連を導くことが可能な値である本手法の再現率の期待値は、約 21%という値になる。前段で述べた、抽出する関連の種類を増やす改善を行った場合、この値は約 15%向上した。

解析エンジンの再現率には、ケース作成精度よりもクラスタ作成精度が低いことが大きく影響していると考えられ、クラスタ作成精度を向上させることで、関連ドキュメントの再現率が向上すると考えられる。今回収集した履歴の分析から、クラスタのルールを何点か改善することで、クラスタの作成精度を 10%程度向上させることが出来ると考えられる。これにより、前段で述べた抽出する関連の種類を増やした改善をあわせて、解析エンジンの再現率を 20%程度向上させることが出来ると見込まれる。

4．まとめ

本手法の特徴の一つとして、精度結果の適合率が 80%という数値からも分かるように、提示される情報にノイズが少ないことが挙げられる。再現率を重視して適合率が低くなることの多い全文検索と補完的に連携させることで効果的なドキュメント検索が実現できると考える。これは、網羅性の高い全文検索によって対象ドキュメントの絞込みを行い、それを足がかりとして、検索結果にノイズの少ない本手法の特徴を生かし、関連を辿って目的のドキュメントを探し出すという使用方法で効果を発揮できると考える。

その他の特徴として、ドキュメントのコンテンツではなく履歴情報を利用するため、画像，音声，映像など、全文検索で扱いにくかったデータも扱うことが可能であることや、コンテンツに比べて、非常に小さなデータ量で解析を行えるという特徴が挙げられる。また履歴データからは、今回のドキュメント関連だけでなくユーザ間での情報フローのような、様々な観点の関連情報を提示することが可能であるなど、ユーザのドキュメント支援に貢献できるシステム実現が可能であると考えている。

＜参考文献＞

[1] Kurt D. Fenstermacher, Mark Ginsburg:“A Lightweight Framework for Cross-Application User Monitoring”, IEEE Computer, Vol.35,No.3,pp.51-59(2002)