人工知能学会第２種研究会資料
SIG-KST-2012-01-02(2012-07-25)

表構造解析とキーワード抽出で付与したメタデータを複合的に用いた表形式文書検索システムの開発

岡田伊策[1]　齋藤稔[1]　大和裕幸[2]　稗方和夫[2]　三浦慎也[3]

Isaac OKADA[1], Minoru SAITO[1] Hiroyuki YAMATO[2], Kazuo HIEKATA[2], and Shinya MIURA[3]

[1]富士通株式会社共通技術本部ナレッジ推進統括部
[1]System Engineering Knowledge Improvement div.,
SYSTEM ENGINEERING TECHNOLOGY UNIT, FUJITSU LIMITED
[2]東京大学大学院新領域創成科学研究科
[2]Graduate School of Frontier Sciences, THE UNIVERSITY OF TOKYO.
[3]東京大学工学部システム創成学科
[3]Faculty of Engineering, THE UNIVERSITY OF TOKYO

アブストラクト:
１シートが１文書に相当し、複数のシートから成る表形式ファイル群を全文検索すると、当該検索語を含んだものが多数抽出される。これら複数の表形式ファイルは都度開かないと、目的文書が確認できない。精度高く検索するには、業務経験による鑑が必要となる。経験に関わらず、より高い精度で目的文書に到達できることが望まれる。
本研究では、表構造の接点情報を解析・行列化して雛形文書行列との類似度でメタデータを付与、かつキーワード抽出した情報もメタデータとして複合的に用いて、目的文書をより高い精度で検索できるようにした。

1. 背景

富士通のシステムエンジニアリング文書は、表形式ファイルであることが多い。１文書が１シートに相当し、それら複数シートからなる１表形式ファイルとしてＤＢに格納される。

富士通のシステムエンジニアリング開発は標準化(SDEM®：エスデム)され、各工程で使われる文書も標準化されている。それらの文書は図 １のような一定の雛形に基づき定型書式化され、上部に工程や文書種別などの業務上の分類情報を持ったシートとして、単独または複数で表形式ファイル形式を形成して、ＤＢ蓄積されている。

システムエンジニアリング文書作成時には再利用を目的に過去の類似する文書を探索する必要がある。複数のシートから成る表形式ファイル群を全文検索すると、当該検索語を含んだものが多数抽出される。これら複数の表形式ファイルは都度開かないと、目的文書が確認できない。精度高く検索するには、業務経験による鑑が必要となる。

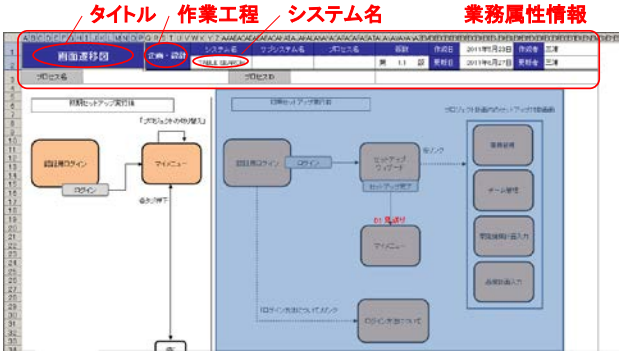

図 １ 表形式文書例

2. 目的

本研究では、富士通のシステムエンジニアリングにおける表形式文書検索で、業務経験に関わらずより高い精度で目的文書に到達できるシステムの開発を目的にした。

具体的には、対象となる表形式文書が特徴として持つテキスト位置や表構造の接点情報を解析・行列化して雛形文書行列との類似度でメタデータを付与した。かつ、キーワード抽出した情報もメタデータとして付与した。

これらを複合的に用いることにより、より高い精度での目的文書到達を目指した。

3. 関連研究

企業内の表形式文書を対象とした既存研究はいくつか存在する[1][2][4]。特に、田中ら[3]は表形式文書の表構造認識手法として、図 2 に示すように表内の罫線が接続、交差する点である節点を用いた手法を提案し、表形式ドキュメントの自動分類システムの提案を行った。

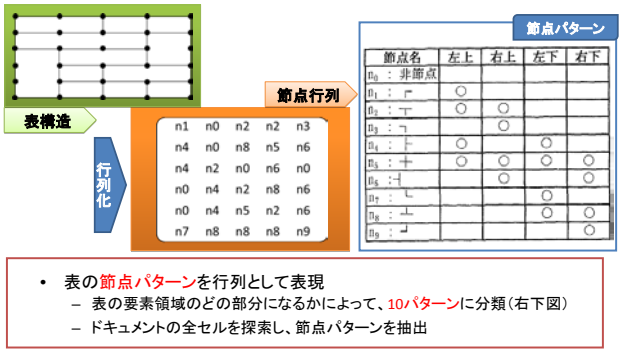

図 2 表構造の抽出手法

しかし、これらの研究では主に表構造文書画像を対象としており、本研究が対象とする表形式ファイルへの適用例や表文書探索への実務応用例も存在しない。

本研究では、システムエンジニアリングにおける表形式ファイルを対象とし、さらに表構造認識技術によるメタデータ付与と、キーワード抽出の複合的に用いる点が従来研究との相違点である。

4. 提案システム概要

図 3 に提案システムの概要を示す。本システムでは、利用者が文書作成時に用いる雛形表形式文書を検索クエリーとして、その業務上の分類メタデータ（以下、「業務属性メタデータ」）と、サンプルテキストの情報（以下、「キーワードメタデータ」）を複合的に用いて、過去の類似する文書を検索し、提示する。

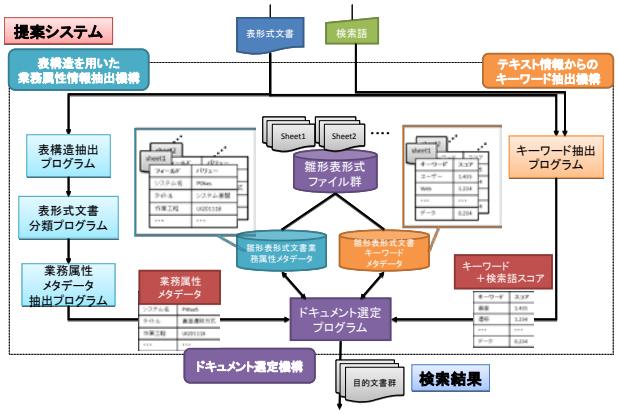

図 3 提案システム概要図

4.1 「業務属性メタデータ」の抽出

文書の「業務属性メタデータ」の抽出では、まず文書を書式により分類し、書式ごとに業務属性情報が存在するテキスト領域に応じた抽出ルールに基づいて、「業務属性メタデータ」抽出を行う。文書分類には、節点行列によって表構造を表現し、それと書式ごとに与えるサンプル文書との行列の類似度を用いる。例えば、図 4 のような業務属性記入欄を持つ文書が存在する場合は、矢印のような抽出ルールを与えることで、文書タイトル、作業工程、システム名、サブシステム名の「業務属性メタデータ」が抽出される。

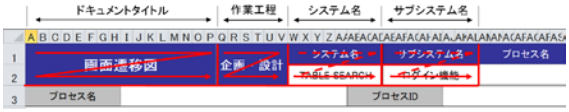

図 4 「業務属性メタデータ」抽出例

表構造の類似度 $Sim_{Tq,Tt}$ は、Tq を雛形文書クエリーの行列化した接点情報、Tt を探索対象表形式文書の行列化した接点情報として、両者のハミング距離 $HD_{Tq,Tt}$ とクエリーの節点数 Nq を考慮して、以下の通り算出した。

$$Sim_{Tq,Tt} = 1 - \frac{HD_{Tq,Tt}}{Nq} \quad (1)$$

また抽出された『業務属性メタデータ』の類似度 $Sim_{Pq,Pt}$ は、以下の通りの算出式とした。

$$Sim_{Pq,Pt} = 1 - \frac{\text{一致する業務属性メタデータ数}}{\text{抽出した全業務属性メタデータフィールド数}} \quad (2)$$

「業務属性メタデータ」の類似度閾値は、0.6 と設定した。

4.2 「キーワードメタデータ」の抽出

表形式文書の「キーワードメタデータ」の抽出は、以下の手順で行った。

① テキスト情報の抽出
・ドキュメント内の全セルを探索
・セルごとにテキストを抽出
・セルごとに改行処理
（隣接セル内テキストの連接防止）
②形態素解析
・MeCab による解析
・不要語除去、名詞抽出
② tf-idf 法によるスコア付け
・検索対象の全ドキュメントを対象に文書 d に

おける単語 t の出現頻度を $tf_{t,d}$、単語の t の逆文書頻度を idf_t として、$tfidf_{t,d}$ は次のように求まる。

$$f_{t,d} = \frac{\text{文書d中における単語tの出現回数}}{\text{文書dにおける全単語数}} \quad (4)$$

$$idf_t = 1 + \ln\left(\frac{\text{全文書数}}{\text{単語tが出現する文書数}}\right) \quad (5)$$

$$tfidf_{t,d} = tf_{t,d} \times idf_t \quad (6)$$

④キーワードを抽出
・スコア上位 10 語を抽出
但し検索語がある場合：スコア上位 9 語+検索語（スコア 1.0）

抽出されたキーワードは、雛型表形式文書のキーワードメタデータと検索語の類似を、コサイン類似度で算出する。

雛形表形式文書のキーワードスコアを Kq、目的文書のキーワードスコアを Kt、ベクトルの内積を $Kq \cdot Kt$、ベクトルの大きさをそれぞれ $|Kq|$、$|Kt|$ とすると、雛形表形式文書と目的文書の類似度 $Sim_{Kg,Kt}$ は以下となる。

$$Sim_{Kq,Kt} = \frac{Kq \cdot Kt}{|Kq| \times |Kt|} \quad (7)$$

「キーワードメタデータ」の類似度閾値は、0.5 と設定した。

4.3 表形式文書の選定

まず、検索対象となる全過去表形式文書に対して、4.1 の手法で表形式文書の分類と「業務属性メタデータ」の抽出を行い、利用者が入力した雛形表形式文書の「業務属性メタデータ」との類似度から文書を選定する。

さらに、「業務属性メタデータ」により選定された過去表形式文書群に対して 4.2 の手法で「キーワードメタデータ」抽出を行い、雛形表形式文書の文書ベクトルとのコサイン類似度からさらに文書を絞り込み最終的な検索結果文書として選定する。

5. 実験と評価

実際の社内の表形式文書群を対象に提案システムの検索性能を実験によって評価した。

5.1 実験概要

表 1 に示すように、社内の 2 つの開発プロジェクトの表形式ファイル約 2,500 ファイルを対象に全 205 パターンの雛形表形式文書をクエリーとしてそれぞれ検索を行った。

表 1 実験対象データ

	表形式ファイル数（Excel形式数）	表形式文書数（シート数）
Project①	694	1,732
Project②	318	857

5.2 評価

(1)「業務属性メタデータ」のみによる検索性能評価

「業務属性メタデータ」のみによる検索性能は、Precision 平均値 0.44、Recall 平均値 0.82、F 値 0.57 となった。

Precision（適合率）P、Recall（再現率）R、F 値 F として、以下のように算出した。

$$F = \frac{2}{1/P + 1/R} \quad (8)$$

比較実験として、全文検索エンジンを用いた実験を行った結果、Precision 平均値 0.18、Recall 平均値 0.99、F 値 0.32 であり、本システムの有効性が示された。図 5 は各検索結果の Precision と Recall を示している。

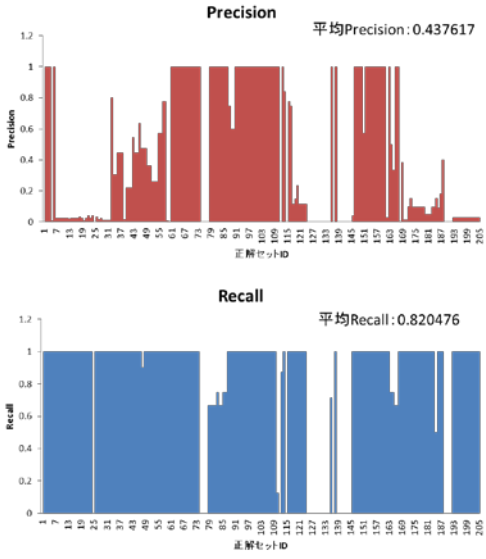

図 5 「業務属性メタデータ」検索性能評価

具体的には、「業務属性メタデータ」により、目的検索文書が、富士通のシステムエンジニアリング開発標準 SDEM®のどのプロセス、どの工程、どのカテゴリに相当するかが特定できた。

しかし、書式が作成者によって編集されてしまっている場合、表形式文書が正しく分類されず、

「業務属性メタデータ」が抽出されないものが存在した。このような表形式文書に対しては節点のパターンだけではなく、セルの距離や行の幅など新たな特徴量を用いる必要がある。

(2) 「キーワードメタデータ」による絞り込み検索性能評価

(1)の検索のうち、特に Precision の低かった 39 パターンの検索結果について「キーワードメタデータ」による絞り込み検索を行った。その結果、39 パターンについて Precision 平均値が 0.0212 から 0.531 まで改善された。図 6 は、絞り込み前後の検索結果の Precision の変化を示している。これにより、「業務属性メタデータ」により検索された結果をさらに「キーワードメタデータ」を用いた絞り込み検索を行うことによる検索性能の向上が確認された。一方で、Recall は 3.2%低下した。

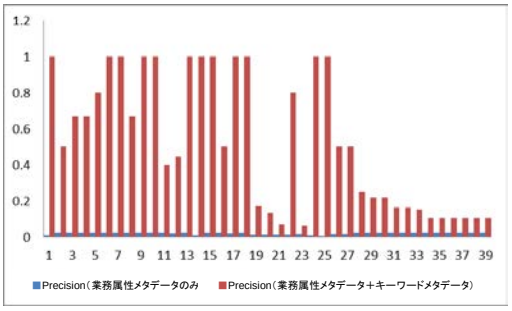

図 6 「業務属性メタデータ」と「キーワードメタデータ」の組み合わせによる絞り込み検索性能評価

具体的には、「業務属性メタデータ」と「キーワードメタデータ」の組み合わせにより、同一業務属性（工程、カテゴリ）配下の複数候補文書群から、目的文書を一層絞り込めるようになった。

しかし、表形式文書内でテキストボックスや図などを多用し、その中にテキスト情報が含まれている場合、テキスト情報の抽出ができず、「キーワードメタデータ」の抽出精度が下がるケースが存在した。そのため、テキストボックスやその他の図内のテキストにも対応したシステムへの改良が求められる。

6. 結論

表構造解析とキーワード抽出で付与したメタデータを複合的に用いた表形式文書検索システムを開発した。具体的には、「業務属性メタデータ」情報による検索と「キーワードメタデータ」による検索を二段階で行うことにより、精度の高い検索システムを目指した。実験では、実際にシステムエンジニアリングで使用されている表形式文書を対象とし、検索性能の評価を行った。

その結果、従来は、絞り込みが不十分で、人手でファイルを開いて黙視確認して特定せざるを得なかった候補量を、絞り込んで削減、本システムの有効性が確認された。

本研究では表形式文書作成時の類似過去表形式文書を探索する場面を想定し、従来の検索システムよりも効率的に表形式文書検索システムを開発した。

なお、表形式文書作成業務の作成効率の向上効果など直接的な有効性については議論されていない。今後は、表形式文書作成支援という観点で議論するとともに、多様な表形式文書作成業務に対応できるような汎用性を実現したい。

参 考 文 献

[1] 安藤智，澤邉一秀，松岡誠，上田弓子，重永信一，“ビジネス文書作成問題における誤り抽出方法，” 情報処理学会研究報告．自然言語処理研究会報告，vol．95，no．27，pp．31-36，1995.

[2] 土井美和子，福井美佳，山口浩司，竹林洋一，岩井勇，“文書構造抽出技法の開発，” 電子情報通信学会論文誌．D-II，情報・システム，II-情報処理，vol．76，no．9，pp．2042-2052，1993.

[3] 田中通，鶴岡信治，吉川大弘，“D-12-27 節点行列を用いた表形式文書の自動分類システム，” 電子情報通信学会総合大会講演論文集，vol．1999，no．2，p．200，1999.

[4] 駱琴，渡邉豊英，杉江昇，“多種帳票文書の構造認識，” 電子情報通信学会論文誌．D-II，情報・システム，II-情報処理，vol．76，no．10，pp．2165-2176，1993.