人工知能学会第2種研究会資料
SIG-KST-2012-02-01(2012-11-16)

複数の年次報告書の差分自動検出による各年報告のトピックメタデータ提示システムの開発

岡田伊策[1]　齋藤稔[1]　笈田佳彰[1]　大和裕幸[2]　稗方和夫[2]

Isaac OKADA[1], Minoru SAITO[1] Yoshiaki OIDA[1], Hiroyuki YAMATO[2] and Kazuo HIEKATA[2]

[1]富士通株式会社 SI 技術サポート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.

アブストラクト

年次報告書のように同一テーマで複数年に渡って毎年書かれる報告書では､各年の報告書の差分理解が重要であり、そのために報告書全体を確認する必要がある。

本研究では、tf-idf の重みづけなどにより各年度の報告書からキーワードを抽出、その変化から各年の差分を自動的に検出して､各報告のトピックをメタデータとして提示するシステムを開発、実際の報告書データに適用してその有効性を示した。また、複合語対応など今後の課題を明確にした。

１．はじめに

特定の終了期限がなく、複数年にまたがって継続的に運営される活動では、一般に、活動固有の定型的な年次報告書が書かれる。当事者や直接・間接的な利害関係を有するステークホルダーにとって、年次報告書は、活動理解のために重要であり、特に各年報告書の差分理解が、進展理解・進捗管理に重要である。このため既存の報告書の全体を確認する必要がある。

しかし、定型書式で書かれているため、各年次報告書の差分を直感的に理解・把握することは人間にとっては難しく、結局既存の報告書を読破して、その差分すなわちトピックを認識することになる。

本研究が課題としたのは、その労力とコスト削減のために、複数の年次報告書の差分を自動検出して各年報告のトピックを提示するシステムの開発である。

複数の年次報告書からキーワードを抽出してメタデータとして付与、それらを基に文書を分類した文書間の差異をトピックワードメタデータとした。

開発した手法は、企業の社外標準化活動の年次報告書に適用して、具体的な効果性を確認すると同時に、トピック詳細の抽出性能などの新たな課題も明確になった。

２．課題解決のためのアプローチ

2.1 提案手法の概要

本研究では、以下の手順を採用した。

（１） 対象とする文書群に、前処理として形態素解析
（２） tf-idf の重み付けなどにより､各文書のキーワードをメタデータとして抽出
（３） 各文書の抽出されたキーワードの Jaccard 係数を使って、文書をクラスタリング
（４） 分類された文書クラスタ内の各文書に特異なキーワードをトピックとしてメタデータ付与。

その結果、各文書のトピックを文書差異として機械的に提示することが可能となった。

2.2 関連研究

文書からの情報抽出に関する研究としては、橋本［１］らによる新聞記事を対象に自動分類し、トピックとなる社会事象を抽出、トピックの構造化により課題発見を可能にする手法の開発や、張［２］らの大規模文書情報源からユーザの意図に合ったレコードを効率的に抽出する手法の提案、川谷［３］の文ベクトルを用いた文書集合間の差異検出法の提案、

などがある。

本研究は、それらの手法を組み合わせて、より高い網羅性でトピック概要を抽出した点が異なる。

2.3 提案手法の手順の詳細

提案手法の手順の詳細は以下の通りである。

(1) 文書の前処理

対象文書に対して、オープンソース形態素解析エンジン MeCab を用いて、名詞のみを抽出。その際、非自立語や数などの不要語を除去した。

(2) キーワードの抽出

「語の共起関係」、「語の共起関係＋idf 法」、「語の連接頻度」の 3 種類のキーワード抽出手法を試行して、各々の抽出精度を比較した。

① 「語の共起関係」適用：
各語の共起する頻度を χ（カイ）二乗値を用いて計算する手法。文書内で既決するため、他の文書における語の出現状況を考慮しないため、他の文書で用いられているような一般語も抽出される可能性がある。

② 「語の共起関係＋idf 法」の組み合わせ適用：
語の共起関係だけでは他の文書での出現回数を考慮しないため、idf 値を算出して出現頻度を算出。χ(カイ)二乗値と idf 値の乗算で重み付けする。

③ 「語の連接頻度」の適用：
語の連接する頻度を算出して、頻度の高い語に重み付けする。

サンプル文書に適用した結果が表 1 の通りである。

「語の共起関係」では「課題」、「議長」、「開催」といった一般語が多く抽出されている。

「語の連接頻度」では「ITU」に関する複合語ばかり抽出されている。また、このサンプル文書には「TDAG 会合」についての記述も多数あり、抽出結果に偏りがあることを示している。

「語の共起関係＋idf 法」は、「語の共起関係」単体に比して、一般語が多く排除された上に、文書内容に関して、幅広くキーワードが抽出されている。

よって、「語の共起関係＋idf 法」が、他の文書での出現頻度や語の重み付けを考慮でき、幅広くキーワード抽出できるとして採用した。

表 1　3 種類のキーワード抽出手法の試行結果

	①語の共起	②共起＋idf法	③語の連接
1	SG	WTDC	ＩＴＵ
2	BDT	BDT	テレコムアジア
3	ＩＴＵ	ＩＴＵ	ＩＴＵテレコムアジア
4	課題	TDAG	ＩＴＵ本部
5	議長	電気通信開発	ＩＴＵテレコム
6	WTDC	ＩＴＵ本部	ＩＴＵ主催
7	開催	ＩＣＴ	ＩＴＵ活動
8	促進	中近東途上	トピックスＩＴＵテレコムアジア
9	電気通信開発	官民パートナーシッププロジェクト	日本ＩＴＵ協会
10	TDAG	自動翻訳システム	ＩＴＵ事務局長

(3) キーワードの抽出

次に、各文書のキーワードの集合(n 個)同士の類似度を Jaccard 係数を用いて分類して、閾値を超える文書を同じ文書クラスタに分類した。

文書クラスタ X, Y について X∪Y の要素を z1, z2,..,zn として，ベクトル **x**=(x1,x2,..,xn)を，xi=1 (if zi∈X), xi=0 (otherwise)として定める.

ベクトル **y** も同様に定めると，Jaccard 係数は次の式で算出される。

$$sim = \frac{\mathbf{x}\cdot\mathbf{y}}{\sum xi + \sum yi - \mathbf{x}\cdot\mathbf{y}} \cdots\cdots\cdots \quad (1)$$

閾値 t を超えるものを、類似文書として同一クラスタに分類することにした。

(4) 各文書のトピックをメタデータとして付与

各文書の各文書の各キーワードの df 値を計算して、閾値 p を超える語をそのクラスタの「テーマワード」として選定することにした。またテーマワード以外の語をその重み順に各文書の「トピックワード」として選定し、両タイプクラスタ内に共通して出現するワードをメタデータとして各文書に付与することにした。概念を図示すると図 1 の通りである。

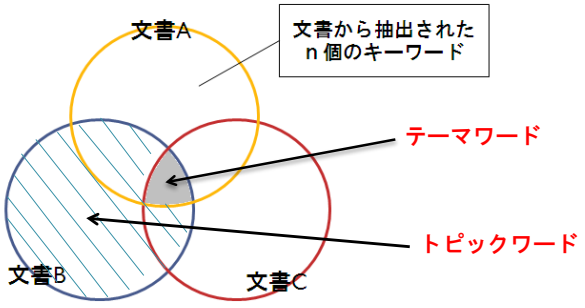


図 1 文書クラスタ内のワードの関係

３．検証

3.1 検証対象データ

提案手法は以下のデータを対象に検証した。

(1)対象文書：ICT 企業の社外標準化活動の年次報告書

(2)対象年度：2003 年度から 2008 年度の 6 年間

(3)対象文書の総数：491 文書データ

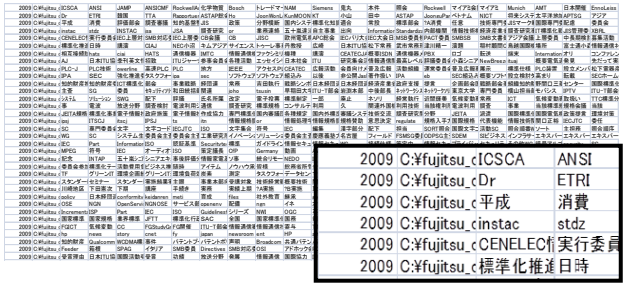
(4)記述内容：定型化された年次報告スタイルをベースに、さまざまな社外標準化活動の活動内容や参加国際会議の内容など

(5)分類パラメータは以下のように設定：

- 文書当たりのキーワード数: 20 語
- Jaccard 係数の閾値 t: 0.25
- df 値の閾値:1.0

対象 491 文書データ全体のキーワード抽出結果は、表 2 の通りである。左 2 列は年度およびファイル名、３列目から抽出されたキーワードのリストを示す。

表 2　全体キーワード抽出結果



抽出したキーワード　の Jaccard 係数を算出して、閾値を設けて分類すると、表 3 のように、164 の文書クラスタに分類された。

表 3 文書分類結果

全文書ファイル数	文書クラスター数	ファイル数別クラスター数				
		2	3	4	5	6
491ファイル	164クラスター	59	49	26	21	9

3.2 有効性検証

ここでは、実際に ITU-D という国際標準化組織に関する 2003 年度から 2008 年度の 6 年間の活動を教師データに取り上げた。

(1) 正解データ

報告書内に存在する「主な活動とトピックス」という項目の内容を報告筆者が意図する正解値とし、この内容と提案手法で抽出された「トピックワード」との一致度合いにより提案手法の有効性を評価した。

実際に報告筆者によって記述されている「トピックス概要」および「トピックス詳細」は、表 4 の通りである。

表 4.　報告書に明示された ITU-D 関連トピックス

年度	トピックス概要	トピックス詳細
2003	•主な活動はSG、TDAG	日本テレコム、東海大学と情報提供・収集
	•世界情報社会サミットに提言	開発途上国のICTプロジェクトのインフラ開発への資金供与の必要性を提言
2004	•主な活動はSG、TDAG	ルーラル通史について、日本のセクターメンバとしてラポーター会合の開催協力
2005	•世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	規制改革のプログラム 、電気通信ネットワーク／技術プログラム 、ｅ－戦略プログラム、料金／財政問題のプログラム 、人材開発 、後発開発途上国（LDC）及び小島嶼開発途上国（SIDS）向け特別プログラム が議題
	•SG定例会議への参画	
2006	•世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	
	•SG定例会議への参画	
2007	•世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	
	•SG定例会議への参画	
2008	•ITUテレコムアジア2008に参加	WiMAX、LTE基地局、パケットオプティカルネットワークのパネル展示
	•SG2会合の開催	自動翻訳の活用、専門家会合運営方法の改善等について、寄書を提出
	•TDAG会合の開催	官民パートナーシップによるプロジェクト（中近東途上国へのワイヤレスプロジェクト）の推進、人材育成の継続、緊急通信への取り組み等の課題に言及
	•WTDC2010開催の準備プロセス	自動翻訳システムを活用し、電子メールで非公式だが円滑な意見交換を提案

(2) 提案手法による実験結果

実験対象全体の 491 文書データを提案手法で処理して、「テーマワード」に「ITU-D」を含む文書クラスタを選択。その文書クラスタの年度別の「トピックワード」は表 5 の通りとなった。

表 5　ITU-D の 6 年分の「トピックワード」

テーマワード	ITU-D、WTDC、ATU-D、電気通信開発、TDAG、実行部門、局長					
トピックワード	2003	2004	2005	2006	2007	2008
	Detailsww	開発途上	開発途上	プログラム	ラポータ	ITU
	開発途上	セクターメンバ	世界電気通信開発会議	ラポータ	SG	ITU本部
	先進	SG	中心メンバー	世界電気通信開発会議	世界電気通信開発会議	ICT
	ICTプロジェクト	中心メンバー	SG	SG	プログラム	中近東途上
	Detailsww	est	プログラム	セクターメンバ	開発途上	官民パートナープロジェクト
	ntt	最終報告	行動計画	行動計画	行動計画	自動翻訳システム
	SG	orporation	援助	途上	途上	電子メール
	中心メンバー	ルーラル通信	運用計画	職員	職員	開発途上
	最終報告	遠隔医療	職員	カタール	セクターメンバ	契約形態
	資金	活動計画	組織ITU	ドーハ	テレセンター	途上
	民間部門	加入	電気通信開発アドバイザリーグループ	テレセンター	プロジェクト実施	ウィルコム
	インフラ開発	Telegraph	研究委員会	プロジェクト実施	回線交換	世界電気通信開発会議
	比較プライオリティー	公共部門	遠隔医療	回線交換	戦略プログラム	アラブ市区事務所

3.3 評価

「正解データ」との一致は、表 6 の通りである。

- 点線で囲んだ部分が「トピックワード」として一致した部分、
- 実線で囲んだ部分が「テーマワード」と一致した部分

である。

なお、資金、プログラムなどの「抽象的な語」の一致は除外した。

表 6　評価（「トピックワード」との一致）

年度	トピックス概要	トピックス詳細
2003	・主な活動はSG、TDAG	日本テレコム、東海大学と情報提供・収集
	・世界情報社会サミットに提言	開発途上国のICTプロジェクトのインフラ開発への資金供与の必要性を提言
2004	・主な活動はSG、TDAG	ルーラル通信について、日本のセクターメンバーとしてラポーター皆皆の開催協力
2005	・世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	規制改革のプログラム、電気通信ネットワーク／技術プログラム、e－戦略プログラム、料金／財政問題のプログラム、人材開発、後発開発途上国（LDC）及び小島嶼開発途上国（SIDS）向け特別プログラム が議題
2006	・SG定例会議への参画	なし
	・世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	2005と同じ
2007	・SG定例会議への参画	なし
	・世界電気通信開発会議開催（2006年3月カタール国ドーハ市）	2005と同じ
2008	・SG定例会議への参画	なし
	・ITUテレコムアジア2008に参加	WiMAX、LTE基地局、パケットオプティカルネットワークのパネル展示
	・SG2会合の開催	自動翻訳の活用、専門家会合運営方法の改善等について、寄書を提出
	・TDAG会合の開催	官民パートナーシッププロジェクトや中近東途上国へのワイヤレスプロジェクトの推進、人材育成の継続、緊急通信への取り組み等の課題に言及
	・WTDC2010開催の準備プロセス	自動翻訳システムを活用し、電子メールで非公式だが円滑な意見交換を提案

実際の活動報告書の「トピックス概要」においては、2003 年度から 2008 年度の全 13 項目中 10 項目（76.9%）において「教師データ（正解データ）」と「トピックワード」との一致がみられた。

「トピックス詳細」においては、2003 年度から 2008 年度で全 10 項目中 4 項目（40%）において「トピックワード」との一致がみられた

3.4 考察

「トピックス概要」については、本提案手法によって、トピックワードとしての抽出が高い網羅性をもってなされた。

「トピックス詳細」については、抽出性能は劣っていた。

提案手法によって選出されなかったトピックとしては、例えば「テレコムアジア 2008」のようにトピックであるにも関わらず、文書内での出現回数が少なかったものや、ＳＧ２（Sub Group 2）のようなあらゆる複合語を作る語があった。後者は「SG2 会合」と「SG2」を別のものとして扱ってしまっているので、語の意味的な同値性を考慮した対策を講じる必要がある。

４．結論

本論の結論としては、各文書から抽出したキーワードをもとに、文書間差異をトピックワードとして検出する手法を提案した。

提案手法を実際にシステムとして実装し、IT 企業の「社外標準化活動年次報告書」という実データを用いて実験を行った

トピックの概要については高い網羅性をもったトピックワード抽出がなされた。

５．今後の展望

トピックワード選定の精度向上のために、2 つのアプローチが考えられる。

キーワード抽出の向上については、語の意味的な考慮を加えた新たな手法の組み合わせを追求することである。

文書クラスタリングと文書間の差異検出については、設定した閾値の最適化や、新たなクラスタリング手法の導入が考えられる。

引き続き、実データを活用して、継続研究していきたい。

参考文献

[1] 橋本 泰一, 村上 浩司, 乾 孝司, 内海 和夫, 石川 正道: 文書クラスタリングによるトピック抽出および課題発見, 社会技術研究論文集, 5, 216-226, （2008 年）

[2] 張建偉,石川佳治,北川博之: トピックを考慮した大規模文書情報源からのレコード抽出情報処理学会論文誌,（2007 年）

[3] 川谷隆彦: 文書集合間の差異検出法と文書分類への応用, 情報処理学会研究報告. 自然言語処理研究会報告, （2002 年）