自 然 言 語 処 理 148－2
（２００２．３．４）

国際会議の論文募集ファイルからのトピックの抽出とクラスタリング

盧 世淼 † 峯 恒憲‡ 雨宮 真人‡

† 九州大学大学院システム情報科学府
‡ 九州大学大学院システム情報科学研究院

〒 816-8580 福岡県春日市春日公園 6 丁目 1 番地
{shimiao, mine, amamiya}@al.is.kyushu-u.ac.jp

我々は，Web ページや電子メールのドキュメントから，ユーザが興味を持つ分野に関する情報だけを選択し，その中に含まれる特定の情報を抽出，クラスタリングを行うことで，抽出した情報間の関連性と，その抽出した情報と，元のドキュメントとの間の関連性を求めることを目的としたシステムの開発を行っている．その一例として，国際会議の論文募集ファイルを対象として，そのファイルからトピックを抽出し，トピック間の関係を求める実験を行ったので，本稿では，その結果について報告する．

キーワード： 情報抽出，情報分類，テキストマイニング，クラスタリング

Topics Extraction and Clustering from Call for Papers Files about International Conference

LU, Shimiao MINE, Tsunenori AMAMIYA, Makoto

Graduate School of Information Science and Electrical Engineering, Kyushu University

6-1 Kasuga-koen, Kasuga, Fukuoka, Japan, 816-8580
{shimiao, mine, amamiya}@al.is.kyushu-u.ac.jp

We are aiming to develop a Web mining system that gathers Web pages, filters them according to user's interests, extracts some entities specified by the users and clusters the entities extracted to discover some relationships between them. In this paper, we propose a Web mining method and discuss preliminary experimental results. The system at first filters the Web pages related to call for papers with SVM(Support Vector Machine) from Web pages gathered through a search engine. Next, it extracts conferences' topics and clusters them. The results clustered are shown through an Explorer like graphical user interface.

Keywords: Information Extraction, Information Filtering, Text Mining, Clustering,

1 はじめに

近年のインターネットの急速な普及は，我々を取り巻く情報環境を大きく変えつつある．特に，World Wide Web(WWW) の普及により，我々は，電子化された膨大な量の情報を容易に入手できるようになった．たとえば，Google (http://www.google.com) で，論文募集に関する情報を調べるため，「call for paper」と入力すると，約2,810,000件もの関連Webサイトを見つけることができる．一方，その膨大さゆえ，全ての情報にアクセスすることは不可能である．検索エンジンは，見つけたドキュメントについて，検索要求文との関連性を基にランキングを行って表示をするが，様々な分野や内容について分類されずに提示されるため，必要な情報を見つけるのに不便な場合が多い．そのため，検索エンジンが返した結果を分類し，図や表などの形式でユーザに提示することで，必要な情報を容易に発見できるようにするための研究が，数多くなされてきた (e.g. [1, 3, 9])．更に，情報の探索だけでなく，膨大な量のドキュメントや情報から，ユーザにとって興味ある情報や知識を発見するマイニング技術も重要度を増してきている．その一分野であるWebマイニングは，Web上の膨大な情報の探索から，特定の情報の抽出や分類，更にはWeb知識ベースの構築にいたる様々な問題を対象としている．

我々は，そのようなWebマイニングの一手法として，特定の情報間の関係や，その情報と，その情報を含むドキュメント間の関係を求めることを目的とした手法の開発を進めている．その一例として，国際会議の論文募集ファイルを対象として，その会議の分野に最も注目されているトピックや，これらトピック間の関連性，更には，そのようなトピックを扱っている会議との間の関係を求めるシステムの開発を行った．このシステムは，文書の自動分類技術や，情報の自動抽出技術，およびクラスタリング技術に基づいて構成されている．

これまで文書の自動分類では，機械学習による分類手法がよく用いられてきた．機械学習に基づく方法では，データ数が大規模な場合や，ユーザの要求によって頻繁に分野が変わるような場合に優れた性能を示している．K-近傍法，決定木，Naive Bayesなどの様々な学習手法が盛んに利用されているが，その中で，Vapnik等によって提案されたSVM(Support Vector Machine)[7, 8] は，新しい統計的学習理論に基づく分類手法として，近年注目を集めている．文献 [4] では，SVMが高次元の入力属性が必要であるようなテキスト自動分類問題において，優れた汎化能力を持つことが述べられている．本研究で開発したシステムでは，このSVMを情報を洗練化するためのフィルタリングの目的で利用している．

文書中の特定の情報を自動抽出する研究は，あらかじめ指定された事柄や実体（例えば，人や場所，日付，製品，価格などの名前）を対象として盛んに行われている．本研究では，抽出対象を国際会議のトピックとして，そのトピックを抽出するために，情報抽出手法としては一般的な，パタンマッチングによる方法を採用した．ただし，トピックの抽出は，なるべく低コストで実現できるように，トピックを表す少数の簡単なパターンと，ドキュメント中でのトピックを含む部分に現れる典型的な表現形式だけを利用することとした．

抽出されたトピックをクラスタリングするために，本研究では，非階層型（non-hierarchical clustering）クラスタリング手法の一つである単連結法 (single link method) を採用した．これは，階層型クラスタリング法に比べて，計算コストが小さく高速であるとの理由からである．

以下，２章では，本研究で開発したシステムの概要について述べた後，SVMを利用した国際会議論文募集ファイルのフィルタリング法，トピック抽出手法，およびトピックのクラスタリング法について述べる．３章では，ユーザインタフェースについて紹介し，４章で，トピックを利用したクラスタリング実験の結果について考察する．

2 システムの構成

本システムの構成を図１に示す．本システムは，主に５つのモジュールから構成されている．

1. 国際会議の論文募集関連のファイルの収集

   ある検索エンジンに対して，検索要求文 (call for paper) を投げ，返された結果から，一定数のファイルを得る．

2. 国際会議の論文募集ファイルの洗練化

   テキスト自動分類技術SVM[8] を用い，先の1で得たファイルのうち，論文募集ファイル

に関連するファイルと，関連性の薄いファイルとに分類する．

3. トピックの抽出

   情報抽出技術を利用し，2 で洗練された論文募集ファイルから，トピックを抽出する．

4. トピックのクラスタリング

   3 で抽出したトピックを各々のクラスタに帰属させ，トピック間の関連性を求める．

5. ユーザインタフェース

   4 で得られたトピック間の関連性，及び，トピックと国際会議間の関連性を提示する．また，ユーザが興味あるトピックについて検索を行い，その関連トピックと共に，提示する．

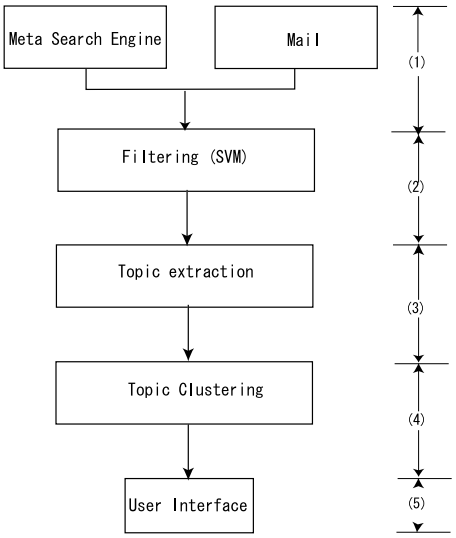

図 1: システムの構成

2.1 SVM を利用した国際会議論文募集ファイルのフィルタリング

検索段階で取得されたファイルの中には，論文募集ファイルと関連性のないファイルも数多く含まれている．そこで，SVM[8] を利用し，収集したファイルから論文募集と関連のないファイルを除去する．SVM を利用した理由は，SVM が高次元の入力属性が必要であるようなテキスト自動分類問題において，優れた汎化能力を持つからである [4].

2.1.1 SVM のアルゴリズム

ここでは，SVM のアルゴリズムについての簡単な説明だけを行う．詳細は，文献 [7, 8, 2] などを参照されたい．

訓練データのベクトルを $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, l$, $y_i \in \{-1, 1\}$, $\mathbf{x}_i \in \Re^d$ とする．

ここで，$\mathbf{x}_i$ はデータ i の特徴ベクトルで，y_i はデータ i が正例 (1) か負例 (-1) かを表すスカラである．線形 SVM では，これらのデータを n 次元 Euclid 空間上で正例と負例に分け，正例と負例の分類境界のマージン（margin）を最大にするような最適な二つの分離超平面 (separating hyperplane) を求める (図２).

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \quad if \ y_i = +1 \tag{1}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \quad if \ y_i = -1 \tag{2}$$

式 (1)，(2) で二つの超平面を定義する．二つの分離超平面の間のマージンは，$\frac{2}{\|\mathbf{w}\|}$ である．すわなち $\|\mathbf{w}\|$ が最小になるような超平面を求める．この問題は Lagrange 乗数を導入して２次最適化問題として扱うことができる．

$$\text{目的関数：} \frac{1}{2}\|\mathbf{w}\|^2 \rightarrow \quad \text{最小化} \tag{3}$$

$$\text{制約条件：} y_i(\mathbf{w} \cdot \mathbf{x}_i) + b) - 1 \geq 0, \ \forall i \tag{4}$$

この問題の解は次の双対問題 (5),(6) を解くことによって得られる．そして，式 (7) により α_i から $\mathbf{w}$ を構築し，マージンを最大にする超平面を求めることができる．

$$\text{目的関数：} \sum_{i=1}^{l} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \rightarrow \text{最大化} \tag{5}$$

$$\text{制約条件：} \sum_{i=1}^{l} \alpha_i y_i = 0, \ \forall i : \alpha_i \geq 0 \tag{6}$$

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{7}$$

また，SVM では式 (5) 中の内積をカーネル関数 $K(\mathbf{x}_i, \mathbf{x}_j)$ で置き換えることによって，非線形データ関数を扱うことができる．カーネル関数には，多項式関数や RBF 関数などもあるが，本システムでは，予備実験の結果，高速な線形カーネル関数を用いることとした．

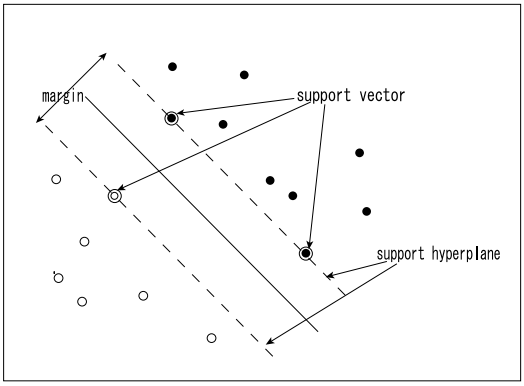

図 2: SVM についての概念説明図

2.1.2 文書の特徴ベクトル

SVM をテキスト分類問題へ適用するために，文書を特徴ベクトルに変換する必要がある．先ず，訓練文書に含まれる全ての単語からなる語彙集合 V をつくると，文書は V に含まれる単語の重みベクトルとして表される[1]．単語の重みとしては，以下の式 (8) で表される TFIDF(Text Frequency Inverse Document Frequencey) 重みを用いる．ここで，$w_{i,j}$ は，文書 i 中の単語 j の重み，$tf_{i,j}$ は，文書 i 中で，単語 j が出現する回数，idf_j は，全文書数 N と，単語 j を含む文書数 n_j との割合 (N/n_j) の対数を取ったものである．

$$w_{i,j} = tf_{i,j} \cdot idf_j \tag{8}$$

次に，訓練文書が目標カテゴリに属する場合，変数 y_i の値を 1 とし，属さない時は-1 として，学習を行う．その後，同様にテスト文書についても特徴ベクトルを作り，SVM で分類を行う．

2.1.3 SVM を利用したフィルタリング実験結果

まず，フィルタリング性能を確かめるための実験を行った．実験には，SVM_light[5, 4] を用いた．また，実験データとしては，Google に検索要求文「call for paper」を与えて得られた Web ページや，DIKU(http://www.diku.dk/research-groups/ topps /Conferences.html) から収集した Web ページを用いた．論文募集ファイルか否かを判断する基準は,「国際会議の名前, トピック，プログラムコミッティー，論文の提出締め切り日，論文の提出に関する情報」を含むこととした．この実験で Kernel 関数として，線形関数を用いた．訓練集合は 604 ファイル（正例 213，負例 391)，テストデータは訓練集合を含まない 253 ファイル（正例 79，負例 174) である．表 1 に，実験結果を示す．

[1] 予備実験では Bi-gram についても試してみたが，差がほとんど見られなかったので，単語ベクトルとした．

Pos.	Neg.	Accuracy	Precision	Recall
79	174	91.07%	84.71%	90.00%

表 1: SVM 実行結果

ここで，Pos.，および Neg. は，それぞれ，正例，および負例の数を表している．Precision は，正例集合に分類された例のうちでの正例であったものの割合，Recall は，正例のうちで，正しく正例集合に分類された割合，Accuracy は，正しく分類された例の割合を示す．この結果，SVM がフィルタリングの目的で有効に機能することが確認できた．

2.2 国際会議の論文募集ファイルからのトピック抽出

膨大な情報の中から，ユーザが興味を持つ項目を抽出して提示することは，必要な情報を選ぶために有効である．ユーザが，特定の会議への論文の応募を決める基準として，その会議の名前や開催時期，場所，締め切り日などがあるが，最も重要なことは，ユーザにとって興味あるトピックをテーマとしてあげているか否かである．そこで，国際会議の論文募集ファイルから抽出する項目として，トピックの抽出を行うこととした．トピックの抽出は，以下の 2 段階で行う: (1) トピックを含むブロック（トピックブロック）の抽出，(2) トピックブロックからのトピックの抽出

2.2.1 トピックブブロックの抽出

トピックブロックを抽出するために，今回は，トピックが箇条書き形式で記述されていると仮定し，また，トピックブロックの開始前に頻出するキーワードの組（例えば,'topics" や"area"，“issue"，“interest"，“theme"）をトピックを抽出するための目印（トピック標識）として，それぞれのパタンを作成した．箇条書き形式は，ファイル形式により異なる．HTML ファイルの場合には，HTML の項目タグのリスト形状で表され，テキストファイルの場合には，マーク形状 (mark shape) が利用される．マーク形状というのは，行頭に連続して出現するスペー

スや” ∗”, ” •”, ” ◦”, ” —” などの記号を指す．そこで，それぞれのファイル形式を考慮したパタンを作成し，利用する．

例えば，図 3 に示した簡単な HTML ファイルからトピックブロックを抽出する場合，まず，トピック標識を探し，その固有表現の後に出現する $('\langle ol\rangle', '\langle \backslash ol\rangle'), ('\langle ul\rangle', '\langle \backslash ul\rangle')$ などの HTML のタグリストの組を切り出す．次に，切り出されたブロックの中から，タグを削除した内容をトピックブロックとして抽出する．

```
<HTML>
……
<H4>The following is a partial list of topics of
interest:</H4></TD>
<UL>
<LI>Hypertext and hypermedia
<LI>Web accessibility
<LI>Intelligent agents
<LI>Resource management
<LI>http and beyond
<LI>Performance and reliability
<LI>Real-time multimedia suppor
</UL>
……
</HTML>
```

図 3: HTML ファイル

また，テキストファイルから抽出する場合には，タグリストの変わりにマーク形状の部分を切り出す．図 4 にテキストファイル中に含まれるトピックリストの例を示す．

```
Topics of interest, but not limited to:

   * All aspects of NLP in agents, e.g. including speech
     processing, language understanding, language production
     and generation, text generation.
   * Evolution of language in embodied systems.
   * Human/Agent communication.
   * Inter agent communication.
……
```

図 4: テキストファイル

2.2.2 トピックの抽出

トッピクブロックからトピックを抽出する際には，句読点記号や接続詞 (例えば “and”, “or”)，および前置詞 (例えば “for”, “from”) をトピックの区切りとして利用した．ついで，抽出されたトピックの中から “stop word” を削除し，単語の複数形を単数形に変換した．このようにして得られた語や句をトピックとした．

2.3 トピックのクラスタリング

トピック間の関係，およびトピックと国際会議との間の関係を求めるため，抽出されたトピックについてクラスタリングを行った．そのクラスタリングの手続きは，大きくわけて，以下の 3 つからなる．

1. 各トピック間の類似度の計算．
2. 1 の類似度計算結果に基づいたベースクラスタの作成．ここでベースクラスタとは，あるトピックと，そのトピックに関連するトピックの集合である．
3. 共通のトピックを持つベースクラスタの結合．

2.3.1 トピック間の類似度の計算

各トピックが表れる文書集合の重複度によって，トピック間の類似度を計算する．トピック間の類似度計算法は，文献 [9] のベースクラスタの類似度計算法を参考にした．

まず，二つのトピック m,n が表れる文書集合をそれぞれ T_m，T_n とし，文書集合のサイズを，それぞれ $|T_m|$，$|T_n|$ とする．また，$|T_m \cap T_n|$ は T_m と T_n の共通文書の数を表す．二つのトピック間の類似度は，以下の条件のどちらかを満たせば 1 とし，そうでなければ，類似度を 0 とする．

$$|T_m \cap T_n|/|T_m| > TH_m \tag{9}$$

$$|T_m \cap T_n|/|T_n| > TH_n \tag{10}$$

ただし，TH_m および TH_n は，実験によって決まる値とする．

2.3.2 ベースクラスタの作成と結合

あるトピックに対して，類似度が 1 となる全てのトピック（関連トピックと呼ぶ）からなるトピック集合を，そのトピックのベースクラスタと呼ぶ．まず，全てのトピック間の類似度を求め，全てのトピックについて，ベースクラスタを作成する．次に，ベースクラスタ間の関連性を求めるために，共通のトピックを持つ複数のベースクラスタを結合する．このクラスタリングアルゴリズムは，単連結法 (single link method) と呼ばれるクラスタリング手法に相当する．

図 5 は，２つのベースクラスタから，１つのクラスタが形成される様子を表している．小円が 1 つのトピックを表し，複数の小円を含む大きな円が，

一つのベースクラスタを表している．２つのベースクラスタ中で破線で結合されている３つのトピックが，それぞれのベースクラスタの共通トピックであり，その共通トピックの結合を通じて，２つのベースクラスタから，１つのクラスタが構成される．

この方法では，一つのベースクラスタが，あるトピックに関連した概念を表し，複数のベースクラスタを結合して形成されるクラスタが，ベースクラスタが表す概念の，上位概念（例えば分野など）を表していると見ることができる．

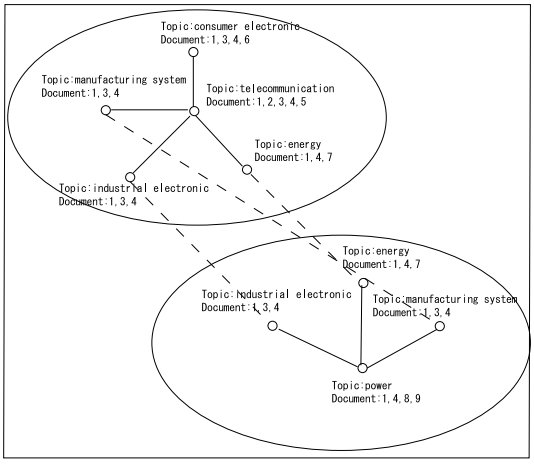

図 5: クラスタリングの例

3　ユーザインタフェース

クラスタリング結果の提示と，検索機能を提供するユーザインタフェースは，以下の３つの部分から構成される．

1. 検索機能部.

   入力されたキーワードに対応する木構造中の節（トピック）の位置を示し，そのトピックに関連した論文募集ファイルのリンクを，関連リンク表示部で表示する．

2. クラスタの木構造表示機能部

   クラスタとクラスタに含まれるトピック, および，各トピックとそのトピックとの関連トピックを，木状に表示する. その木の根は，抽出したトピックの全体を表し, 根の直下の節が各クラスタを表す．各クラスタに含まれるトピックが，クラスタの下位の節として表示され, 更にそのトピックの関連トピックが，木の葉として表される．

3. 関連リンク表示部.

   木構造表示部で表示されるトピックを持つ国際会議の論文募集ファイルへのリンクとその論文募集ファイルから抽出されたトピックを表示する部分である. 論文募集ファイルへのリンクをクリックすると，ファイルの内容を表示するためのウインドーが開く．

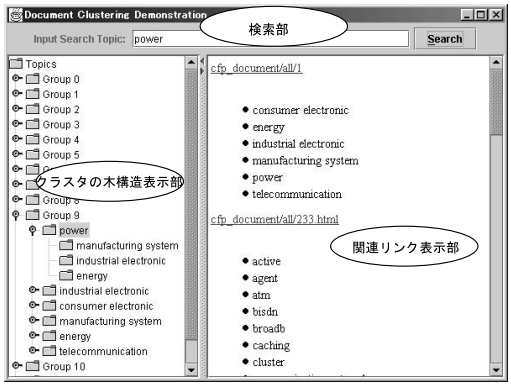

図 6: ユーザインタフェース

4　実験と考察

4.1　トピック抽出実験

まず，実験に用いる Web ページの取得のため，Goo(www.goo.ne.jp) に検索要求文「call for paper」を与え，収集できた Web ページ 951 個に対して，SVM^{light} でフィルタリングを行った．その結果，485 個のファイルが正例とした分類された．つぎに，この 485 個のファイルに対して，トピック抽出を行った．その結果を表 2 に示す. トピックを抽出できたファイル数は全体の 56.5% と半数を超えた程度にすぎなかったが，箇条書き形式でトピックが記述されていたにもかかわらず，抽出に失敗したファイル数はわずか 18 であった．一方，抽出すべき対象ファイルの 93.8%についてトピックブロックを抽出し，更に，そのうちで抽出されたトピックの 79.2%が正確に抽出され，部分誤りも含めると抽出精度は約 88.7%となることを考えると，本システムのトピック抽出性能は，低くは無いと言える．

4.2　クラスタリング実験の手順

クラスタリング実験には，表 5 で示したように，トピックを抽出できた 274 個の Web ページファイルから得たトピックに加え，124 個の論文募集ファ

表 2: トピック抽出実験の結果

実験ファイル数	抽出された		抽出できない
485	274		214
抽出されたファイル数	正しい	誤り	部分誤り
274	217	31	26

表 3: トピックの出現頻度と，その出現頻度を持つトピックの数

トピック出現頻度	トピック数	トピック出現頻度	トピック数
2	488	5	56
3	197	6	45
4	93	6 以上	90

イルに関連する電子メールファイルから得たトピックを用いた．表 3 に，トピックの論文募集ファイル中での出現頻度と，その出現頻度に対応するトピックの数の関係を示す．この結果，できるだけ多くのトピックを利用するため，二つ以上の文書で出現するトピックを用いることにした．その時のトピック数は 969 であった．実験は，以下の手順で行った．

1. 収集された論文募集ファイルのトピックを抽出する．

2. 抽出されたトピックから，そのトピックに関連したトピックの集合をもとめ，そのトピックのベースクラスタを作成する. このとき，第 2.3.1 節で述べたトピック間の類似度でのパラメータ TH_m および TH_n は，それぞれ 0.5 とした．ここで，関連トピックを持たないトピックは，特別なクラスタに入れた．

3. 共通のトピックを持つベースクラスタ同士を結合し，クラスタを生成する．また，トピックを二つしか含まないベースクラスタ中のトピックも，関連トピックがないトピックと同様に扱って, 特別なクラスタに入れる．

4.3 クラスタリング結果の分析

図 5 に，クラスタリング結果を示す．図中のトピック例に示されているトピックは，そのクラスタ中に含まれるベースクラスタの中心となるトピックのうち，関連トピックの数が多いトピック[2]である．また，トピック例の示されていない 2 番のクラスタは，関連トピックがないトピックを扱う特別なクラスタを表している．

[2]つまり，その分野において，注目されているトピックと見ることができる．

この実験で使った国際会議の論文募集ファイルのうちのほとんどが計算機科学関係であったため，図 5 からわかるように，得られたトピックのほとんどが計算機科学分野のクラスタ（グループ 0) に分類されている．その一方，少数ではあるが，検査，生物化学，電子，電気の各々の分野のトピックが，3, 6, 7, 10 番の各々のクラスタに，はっきりと分けられていることも表 4 は示している．

表 5: クラスタリングの結果

クラスタ番号	トピック数	クラスタ番号	トピック数
0	603	7	3
1	3	8	3
2	326	9	6
3	3	10	3
4	4	11	3
5	3	12	3
6	8		

また，表 6 は，表 4 中に含まれるベースクラスタの中心となるトピックと，その関連トピックの例を表している．その表を見てみると, トピック network architecture は,security, protocol, service, wireless などのトピックと関連性が高いことが分かる. 同様に, トピック natural language interface は，information retrieval, ontology, machine translation, information extraction などと関連性が高いことが示されており，直感に合う結果が得られていることがわかる．さらに, 例として”fault modeling” トピックから,“IEEE European Test Workshop” と”1st IEEE Latin-American Test Workshop” という名前の論文募集ファイルが検索された. これらの会議の分野では，test, reliability, defect, thermal testing, simulation, process control などのトピックが，最も注目されていることも分かる．

5 おわりに

本稿では，国際会議の論文募集ファイルを対象にして，国際会議のトピックを抽出し，その抽出されたトピックをクラスタリングすることで，抽出されたトピック間の関係や，トピックと国際会議との関係を示す方法を提案し，その実験結果について議論した．実験に利用したファイルのほとんどが，計算機科学分野の国際会議に関するものであったため，作成されたクラスタの要素数には片寄りが見られたが，計算機科学以外の分野についても明瞭に分類できていることがわかった．また，あるトピック

表 4: クラスタの例

クラスタ番号	トピック例	クラスタ番号	トピック例
0	electronic commerce, telecommunication, tracking, distributed database, information retrieval	7	power, industrial electronic, energy, consumer electronic, manufacturing system
1	corba, java, dcom	8	maj, theme, future of
2		9	panel, presenter, tutorial
3	defect, thermal testing, fault modeling	10	magnetic, proces, monitoring
4	hypertext, text corpora, statistical model	11	interpretation, composition, artificial
5	movement, film, media	12	power electronic, modelling, diagnostic
6	biological, psychological, lo angele, evolution, chemical, ucla		

表 6: 関連トピックの例

network architecture	natural language interface	magnetic	java
security network protocol service wireless	information retrieval ontology machine translation information extraction requirement engineering	process monitoring	dcom corba
pattern recognition	**parallel application**	**electronic commerce**	**decision tree**
multimedia image process neural network computer vision image analysis	distributed digital library agent mobile computing the internet multimedia technology internet service domain specific language collaboration technology	security browser replication payment virtual marketplace rational information agent authorization auction	planning neural network reinforcement learning problem solving

と関連するトピックの数によって，その分野で最も注目されているトピックについても示すことを試みたが，これについては更に分析が必要である．またトピック間の類似度値についても詳細に設定することで，更に詳細なトピック階層を作成することも可能となろう．今後は，更に実験対象のファイル数を増やし，これらの不十分な点について研究を行うとともに，先に提案した国際会議情報の管理システム [6] に本手法を適応する予定である．

参考文献

[1] Robert B. Allen, Pascal Obry, and Michael Littman. An interface for navigating clustered document sets returned by queries. In *Proceedings of the ACM Conference on Organizational Computing Systems(COOCS)*, 1993.

[2] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR conference on Research and development in information retrieval*, pages 318–329, 1992.

[4] T. Joachims. Text categorization with support vector machine:learning with many relevant features. In *Proceedings of Tenth European Conference on Machine Learning(ECML-98)*, pages 137–142, 1999.

[5] Thorsten Joachims. Svm light, http://svmlight.joachims.org.

[6] Tsunenori Mine, Makoto Amamiya, and Teruko Mitamura. Conference information management system : Towards a personal assistant system. In *Proceedings of The First Asia-Pacific Conference on Web Intelligence (WI-2001)*, pages 247–253, 10 2001.

[7] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[8] Vladimir N. Vapnik. *Statistical Learning Theory*. JOHN WILEY & SONS, INC., 1998.

[9] Oren Zamir and Oren Etzioni. Web document clustering:a feasiblity demonstration. In *Proceedings of the 21th Intl. ACM SIGIR Conference*, pages 46–54, 1998.