グラフ的手法による国際会議プログラム情報の解析

A Graph-Based Analysis of Conference Programs

野呂 智哉 根岸 秀典 徳田 雄洋

Tomoya NORO Hidenori NEGISHI Takehiro TOKUDA

東京工業大学 大学院情報理工学研究科

Graduate School of Information Science and Engineering,
Tokyo Institute of Technology

{noro,negishi,tokuda}@tt.cs.titech.ac.jp

現在，世界中で数多くの国際会議が開催されており，その全体的連関を把握することは困難である．一方，インターネットの普及により，国際会議に関する情報 (会議名，発表論文題目，発表者名等) を Web 上から容易に入手することが可能となっている．これらの情報を取得して各国際会議間の関係をグラフで表現し，各国際会議間の相互連関，類似点や相違点，その国際会議における主要な話題や特定の話題に関する中心的な国際会議を発見する手法について提案する．

1 はじめに

近年，世界中で数多くの国際会議が開催されている．その分野は多岐に渡り，歴史のある有名な会議だけでなく最近発足した新しい会議も多く，個人で把握できる範囲は限られている．特に新しい境界領域も扱う会議の場合，以下に示す状況にしばしば遭遇する．

- 類似したテーマについて，異分野の会議がそれぞれ関連して，または，独立に議論を行っている．
- 境界領域の研究成果は，どの会議で発表すべきか判断が難しい．

例えば，セマンティック Web に関する研究は，ISWC (International Semantic Web Conference) [1] における中心的議題であるが，Web 全般を扱う WWW (International World Wide Web Conference) [2]，Web 工学を扱う ICWE (International Conference on Web Engineering) [3]，データベースを扱う VLDB (International Conference on Very Large Data Bases) [4] でも議論がなされている．さらに，LREC (International Conference on Language Resources and Evaluation) [5] という自然言語処理 (特に言語資源の構築や利用) に関するテーマを扱う会議でも，セマンティック Web に関する研究が発表されている．しかし，同じ自然言語処理分野の会議であっても，COLING (International Conference on Computational Linguistics) [6] では，セマンティック Web に関する議論は活発ではない．このような会議間の関係を知ることは，そのテーマにおける現状を把握し，研究領域の全体像を把握する上で有益であるが，容易ではない．

一方，現在では，国際会議に関する情報 (開催期間，開催地，論文募集要項，発表論文題目と著者のリスト，委員のリスト等) は，Web 上で公開されることが多い．これらの情報は無料で入手することができ，開催後も削除されずに残っていることも少なくない．これらを自動的に取得し，解析することにより，その会議で扱われているテーマや，会議間の関係を発見することができると考えられる．

本研究では，Web 上で公開されている国際会議に関する情報をもとに国際会議のネットワークを構築し，そのネットワークから，国際会議間の関係や，その会議で扱う主要な話題，特定の話題に関する中心的な会議等を発見する手法について提案する[7]．

[1] ISWC2004 — http://iswc2004.semanticweb.org/
[2] WWW2005 — http://www2005.org/
[3] ICWE2005 — http://www.icwe2005.org/
[4] VLDB2005 — http://www.vldb2005.org/
[5] LREC2004 — http://www.lrec-conf.org/lrec2004/
[6] COLING2004 — http://www.issco.unige.ch/coling2004/ (プログラム等の情報は既に削除されている)
[7] 国際会議に関する情報を Web から自動取得する手法については今後の課題とし，本論文では手動で収集する．

2 分析に利用する国際会議情報

国際会議に関する情報として，以下のような情報が挙げられる．

基本情報: 正式名称，略称，開催期間，開催地，Web ページ URL

論文募集要項: 背景，目的，議題

発表プログラム: 論文題目，著者 (名前，所属)，セッション名

委員: 議長 (名前，所属)，プログラム委員 (名前，所属)

併設ワークショップ: 正式名称，略称

一方，分析に利用する国際会議情報は，以下の 3 つの条件を満たしていることが望ましい．

1. 国際会議間の関係を分析する上で有用な情報である．
2. 計算機による処理がしやすい．
3. 大多数の国際会議において，その情報を Web から無料で取得できる．

以上を踏まえ，本研究では，以下の情報を利用する．

発表プログラム: 論文題目，著者 (名前)，セッション名

委員: 議長 (名前)，プログラム委員 (名前)

発表プログラムの情報は大多数の会議で取得可能であり，分析対象としてふさわしい．しかし，会議によっては，論文題目と発表者名だけが公開され，発表者の所属は不明であることも多いため，発表者の所属に関する情報は除外する．

議長，プログラム委員等の情報は，会議の内容を直接表すものではない．しかし，1 つの会議の発表者やプログラム委員が，もう 1 つの会議の発表者やプログラム委員である場合，これら 2 つの会議の間には何らかの関係があると考えられるため，論文発表者と同様に扱う．

論文募集要項は，その会議の内容を表す情報を多く含むが，その書式や内容，量は会議によって様々である．例えば，ある会議では，扱う議題を箇条書きで簡潔に示しているのに対し，普通の文章で詳細に述べている会議もある．これらを計算機で処理することは困難であり，今回の分析対象から除外した．

大規模な会議に併設するワークショップが扱う議題は，その国際会議の議題と類似している (もしくは，その中の 1 分野に特化している) ことが一般的であり，その間には非常に強い関係があることが推測できるが，今回は分析の対象から除外する．

3 国際会議情報の分析

国際会議情報の分析は，以下の手順で行う．

1. 各国際会議の情報から特徴語を抽出する．
2. 抽出した特徴語をもとに 2 つの会議間の関連度を算出し，グラフを作成する．
3. 作成したグラフをもとに会議間の関係を分析する．

3.1 特徴語の抽出

国際会議 X に関する情報の中の各単語 w の出現頻度を算出し，一定の閾値 t より大きいものを特徴語とする．

$$\mathrm{feat}(X) = \{w | \mathrm{tf}_X(w) > t\}$$

ただし，$\mathrm{tf}_X(w)$ は，会議 X の情報中における特徴語 w の出現頻度を表す．また，論文題目とセッション名は TreeTagger[3] で形態素解析し，出力される基本形を特徴語として利用する．

単純に単語の出現頻度で特徴語を決定すると，一般的な単語 (どの分野でも一般的に用いられる単語) が特徴語として抽出されてしまう可能性がある．そこで，特徴語となり得る単語のリストを予め人手で用意しておき，そのリストにない単語は，出現頻度が大きくても抽出しないようにする[8]．

人名についても，同様に抽出する．ただし，人名は姓と名に分割せず，フルネームで 1 語とする．また，人名の場合は，特徴語となり得る単語のリストに相当するものは用意せず，すべてを候補とする．

単純に単語を特徴語として利用することには問題がある．それは，同一の単語が，分野によって異なるものを指すことがあることである．例えば，ISWC において，“semantic” という語は “semantic Web”

[8] IDF 等を利用することにより一般的な語を排除する方法もあるが，IDF を算出するためには大量のデータが必要となる．今回は会議情報を人手で収集しており，IDF の算出に十分な量ではないため，予め人手で単語リストを用意した．

$$\mathrm{rel}(X,Y) = \frac{\sum_{w \in \mathrm{feat}(X) \cap \mathrm{feat}(Y)} \mathrm{tf}_X(w)}{\sum_{w \in \mathrm{feat}(X)} \mathrm{tf}_X(w)} \times \frac{\sum_{w \in \mathrm{feat}(X) \cap \mathrm{feat}(Y)} (\mathrm{tf}_X(w) \times \mathrm{tf}_Y(w))}{\sqrt{\sum_{w \in \mathrm{feat}(X) \cap \mathrm{feat}(Y)} \mathrm{tf}_X(w)^2} \times \sqrt{\sum_{w \in \mathrm{feat}(X) \cap \mathrm{feat}(Y)} \mathrm{tf}_Y(w)^2}}$$

図 1: 会議間の関連度

や “semantic (Web) service” という複合語の一部として使われることが多いが，自然言語処理を扱う会議である COLING や ACL (Annual Meeting of the Association for Computational Linguistics) [9]では，“semantic analysis” や “semantic approach”，“semantic role”，“semantic interpretation” 等の複合語の一部として使われる．これらを区別するためには，1 語ずつ独立に扱うだけでなく，複数語の組み合わせも考慮する必要がある．さらに，LREC では，“corpus” と “annotation” の 2 語を “corpus annotation” のように 2 語が連続する形で使用するだけでなく，“corpus semantic annotation” や “annotation of Japanese corpus” のように単語間に別の語が入ることもある．そこで，同一論文題目中における 2 単語の共起頻度を算出し，一定の閾値を超えるものを特徴語 (対) として利用する．ただし，共起頻度の算出の対象とする語は名詞，動詞，形容詞，副詞とし，2 語のうちの少なくとも 1 語は，前述の特徴語となり得る単語のリストに含まれることとする．

3.2 関連度の計算

直感的に，会議 X が会議 Y と関連している場合，会議 X が扱う議題の多くを会議 Y でも扱っていると考えられる．逆に，会議 X が扱う議題を会議 Y で扱っていなければ，関連性が低いことになる．そこで，2 つの会議 X，Y 間の関連を表す尺度として，関連度を図 1 のように定義する．ただし，$\mathrm{feat}(X) \cap \mathrm{feat}(Y) = \emptyset$ の場合は，0 とする．式の第 2 項は，両方の会議に共通して出現する特徴語のみによるベクトルの余弦であり，それと全特徴語の合計頻度に占める共通の特徴語の合計頻度の割合の積を，会議 X の会議 Y に対する関連度としている．

例えば，国際会議 X，Y から抽出できた特徴語 a，b，c，d，e の出現頻度が表 1 のとおりであった

表 1: 会議 X，Y 中の各特徴語の出現頻度

	a	b	c	d	e
会議 X	10	5	5	0	0
会議 Y	6	3	3	5	3

とする．このとき，関連度は以下のようになる．

$$\mathrm{rel}(X,Y) = \frac{20}{20} \times \frac{10 \times 6 + 5 \times 3 + 5 \times 3}{\sqrt{10^2 + 5^2 + 5^2} \times \sqrt{6^2 + 3^2 + 3^2}} = 1$$

$$\mathrm{rel}(Y,X) = \frac{12}{20} \times \frac{10 \times 6 + 5 \times 3 + 5 \times 3}{\sqrt{10^2 + 5^2 + 5^2} \times \sqrt{6^2 + 3^2 + 3^2}} = 0.6$$

この例より明らかであるが，関連度は非対称である．

$$\mathrm{rel}(X,Y) \neq \mathrm{rel}(Y,X)$$

この関連度を算出することにより，国際会議を各ノードとする有向グラフを作成できる．

情報検索や文書分類において，文書やクエリの類似度を計算する際，余弦を利用することが一般的である．しかし，国際会議間の関連性を算出する際に余弦をそのまま利用することには問題がある．例えば，先述の例のように，会議 X に出現する特徴語の集合が会議 Y に出現する特徴語の集合の部分集合になる場合，会議 X は会議 Y が扱う議題の一部に特化した会議であると考えられ，会議 X は会議 Y と関連があると言える．ところが，余弦による関連度の算出では，この包含関係に対して適切なスコア付けができない．図 1 の定義のように，共通する特徴語のみを対象に余弦を計算し，その特徴語の合計頻度が全体に占める割合との積を求めることにより，上述の問題を解決できると考えられる．

3.3 国際会議の分類

Dorow らは，curvature を利用して名詞を分類する手法を提案している [1]．その手法は，以下のと

[9] ACL2005 — http://www.aclweb.org/acl2005/

おりである[10].

1. コーパスから “and”，“or”，コンマで区切られた名詞対を獲得し，名詞をノードとし，名詞対を辺で結合したグラフを作成する．
2. 各ノードについて，curvature を算出する．
3. curvature が閾値以下であるノードを削除する．

国際会議分類でも，同様の手法を利用できる．ただし，Dorow らの手法は無向グラフを利用しているのに対し，我々の関連度によるグラフは有向グラフであるため，無向グラフに変換する必要がある．そこで，関連度を利用し，2 つの会議間の類似度を以下のように定義する．

$$\mathrm{sim}(X,Y) = \sqrt{\mathrm{rel}(X,Y) \times \mathrm{rel}(Y,X)}$$

この類似度を利用して無向グラフを作成することにより，Dorow らの手法を適用することができる．

3.4 中心的な会議の発見

Erkan らは，グラフを利用し，ドキュメント集合から重要文を抽出する手法 (LexRank) を提案している [2]．この手法は，対象とするドキュメント集合を表すグラフから得られる行列 $\mathbf{B}$ について，以下の式を満たすベクトル $\mathbf{p}$ を求めることにより，重要文を決定する．

$$\mathbf{p} = \mathbf{B}^{\mathrm{T}}\mathbf{p}$$

ただし，行列 $\mathbf{B}$ の要素は，各列で総和が 1 になるように正規化されている．同様の手法を，国際会議のグラフから得られる関連度 (または類似度) を各要素とする行列に適用することにより，ある特定の議題に関して中心的な会議を発見できると考えられる．

LexRank では，予め内容の近いドキュメント集合を用意する必要があるが，この国際会議のグラフでは，先に前節の手法で国際会議を分類し，特定の議題に関する会議の集合を用意することにより，LexRank と同様の手法を適用できるようになる．

4 評価実験

前節の手法により，どの程度会議間の関係を発見できるかを調べるため，小規模な実験を行った．

[10] この手法とは別に，辺をノードとしたグラフを作成して分類する手法も提案している．

4.1 国際会議データの収集と特徴語の抽出

まず第一に，国際会議情報として，以下の 4 会議の発表論文題目，著者名，議長，委員の情報を手動で収集した．

- ISWC2004 (セマンティック Web)
- WWW2005 (Web 全般)
- VLDB2005 (International Conference on Very Large Data Bases) (データベース)
- LREC2004 (言語資源)

収集した論文数，発表者数 (のべ人数)，議長，委員数を表 2 の 2〜4 列目に示す[11]．LREC は非常に大規模な会議であり，発表論文数や発表者数が他の会議に比べて多い．会議の規模の違いが分析結果に影響を与える可能性が考えられるが，それについては後述する．

次に，特徴語の候補となり得る語として 635 語を予め人手でリストアップし，収集した国際会議情報から特徴語を抽出した．ただし，特徴語の抽出において，単語出現頻度，人名頻度，共起頻度の閾値を，それぞれ 0，1，1 とした[12]．すなわち，単語に関しては，1 回でも出現した場合には特徴語として採用し，人名と共起単語対については 2 回以上出現したものを採用する．採用した特徴語の数を表 2 の 5〜7 列目に示す．規模の大きい LREC から抽出できる特徴語が多いことは容易に予想できるが，一方，VLDB から抽出できる単語共起対の数が，ほぼ同じ規模の ISWC より少ないことが分かる[13]．

4.2 関連度と類似度の算出

抽出した特徴語をもとに算出した各国際会議間の関連度を図 2 に示す．ただし，$\mathrm{Rel}_{\mathrm{freq}}$，$\mathrm{Rel}_{\mathrm{cooc}}$，$\mathrm{Rel}_{\mathrm{person}}$ は，それぞれ単語，共起単語対，人名をもとに算出した関連度である．

単語，共起単語対，人名のいずれをもとに算出した関連度でも，ISWC と WWW の間に関連がある (ISWC の WWW に対する関連度の方が若干高い)

[11] 議長や委員の情報については，議長等の上位の人物の情報のみを公開している会議と，各トラック (セッション) を担当する委員まで全て公開している会議がある．後者の場合，その数は 100 人を優に超えることもある．今回は，各トラックごとの委員は対象とせず，上位の主要なメンバーのみを収集した．

[12] 人名は，発表者と議長，委員を区別せずに統合して頻度を算出した．

[13] 1 回のみ出現する共起単語対も含めると，ISWC は 819 個，VLDB は 807 個である．

表 2: 利用した国際会議情報

	発表論文数	発表者数	議長，委員数	単語	共起単語対	人名
ISWC2004	153	504	19	105	105	67
WWW2005	300	842	20	169	146	67
VLDB2005	142	510	30	96	63	50
LREC2004	529	1751	7	289	599	279

$$\mathrm{Rel}_{\mathrm{freq}} = \begin{pmatrix} & \mathrm{ISWC} & \mathrm{WWW} & \mathrm{VLDB} & \mathrm{LREC} \\ \mathrm{ISWC} & 1.000 & 0.620 & 0.103 & 0.188 \\ \mathrm{WWW} & 0.533 & 1.000 & 0.146 & 0.149 \\ \mathrm{VLDB} & 0.134 & 0.169 & 1.000 & 0.161 \\ \mathrm{LREC} & 0.099 & 0.099 & 0.064 & 1.000 \end{pmatrix}$$

$$\mathrm{Rel}_{\mathrm{cooc}} = \begin{pmatrix} & \mathrm{ISWC} & \mathrm{WWW} & \mathrm{VLDB} & \mathrm{LREC} \\ \mathrm{ISWC} & 1.000 & 0.361 & 0.000 & 0.265 \\ \mathrm{WWW} & 0.240 & 1.000 & 0.046 & 0.141 \\ \mathrm{VLDB} & 0.000 & 0.094 & 1.000 & 0.000 \\ \mathrm{LREC} & 0.013 & 0.018 & 0.000 & 1.000 \end{pmatrix}$$

$$\mathrm{Rel}_{\mathrm{person}} = \begin{pmatrix} & \mathrm{ISWC} & \mathrm{WWW} & \mathrm{VLDB} & \mathrm{LREC} \\ \mathrm{ISWC} & 1.000 & 0.083 & 0.000 & 0.012 \\ \mathrm{WWW} & 0.047 & 1.000 & 0.026 & 0.000 \\ \mathrm{VLDB} & 0.000 & 0.035 & 1.000 & 0.000 \\ \mathrm{LREC} & 0.006 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

図 2: 関連度

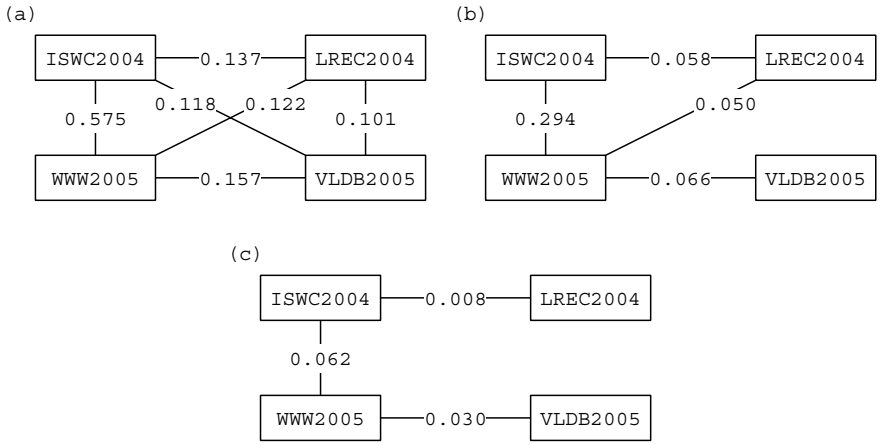

図 3: 類似度

ということが分かる．また，ISWC，WWW，VLDB の 3 会議の LREC に対する関連度が，逆の関連度に比べて高くなっている．これは，LREC の規模が他の 3 会議に比べて大きく，抽出できた単語の数も大きいことが要因であると考えられる．大規模な会議で扱われる議題は，小規模な会議で扱われる議題より多様であることが多く，その意味では，小規模な会議の大規模な会議に対する関連度が高くなることは不思議ではない．しかし，この LREC の例のように，実際の (直感的な) 関連性に比べて過大な値が出る可能性もある．これを防ぐためには，何らかの正規化が必要であると考えられるが，その手法については今後の課題とする．

図 3(a)，(b)，(c) に，それぞれ単語，人名，共起単語対をもとに算出した類似度を示す．これより，関連度の結果と同様，ISWC と WWW の間には共通の議題が多く，類似性が高いことが分かる．一方，関連度では，会議の規模が与える影響が大きかったが，類似度では，双方向の関連度の積 (の平方根) をとるため，影響を小さくすることができる．

curvature による分類はこの類似度を用いて行うが，今回の実験では，扱うデータが少量であるため，その有効性を確認するには至っていない．しかし，ISWC と WWW の間の類似度とそれ以外の類似度の差が十分あることから，有効であると期待している．

4.3 特定の議題に関する中心度の算出

LexRank による中心度の計算は，第 3.4 節で述べたように，国際会議を予め分類し，特定の議題について類似した会議集合に対して行う．今回の実験では，扱うデータが少ないため，会議の分類はしない．その代わりに，特定の議題として，前節の関連度算出の際の特徴語抽出に使用した特徴語候補単語リスト (635 語) から Web，ontology，RDF，OWL，XML の 5 語を選択し，この 5 語について再度算出した関連度をもとに中心度を算出した．各特徴語の出現頻度と出現頻度の高い共起単語対を表 3 に，関連度を図 4 に，中心度を表 4(2，3 行目) に示す (中心度は，最大値が 1 になるように正規化してある)[14]．参考として，関連度ではなく類似度を利用して算出した中心度を表 4(4，5 行目) に示す．

関連度を利用した中心度は，単語をもとにした場合は ISWC が最も高く，共起単語対をもとにした場合は WWW が最も高くなっている．特徴語の抽出に利用した単語リストは 5 語と少ないため，単語の出現頻度の分布に大きな差は生じない．その結果，単語をもとにした場合，ISWC，WWW，LREC の中心度の差は小さい[15]．一方，共起単語対は，単語そのものと比べて多様であるため，出現頻度の分布の差が大きくなり，中心度の差が明確になる．これより，中心度の算出には，単語そのものを利用するより，共起単語対を利用する方が良いと考えられる．

類似度を利用した中心度の算出でも，ほぼ同様の結果が得られる[16]．しかし，中心度の算出には，類似度による無向グラフよりも，関連度による有向グラフを利用する方が良いと考えられる．なぜなら，ある 2 つの会議の間に関連がある場合，有向グラフでは，その 2 つの会議の上下関係 (どちらがより重要，中心的な会議か) を表現できるのに対し，無向グラフでは対等に扱われるからである．実際，共起単語対をもとに算出した類似度を利用した中心度に比べ，共起単語対をもとに算出した関連度を利用した中心度の方が，会議間の差が明確であり，有向グラフであることの効果が見える．

5 おわりに

本研究では，国際会議のプログラムの情報を利用し，国際会議間の関係を分析する手法について考察した．さらに，実験を行ったところ，小規模な実験ではあるが，その有効性を期待できる結果が得られている．

現在，国際会議や学術論文等に関する Web アプリケーションとして，EventSeer[17]，CiteSeer[18]，Google Scholar[19]等がある．本研究は，これらのアプリケーションと競合するものではなく，連携することで，より便利なアプリケーションが実現できると考えられる．

以下に，今後の課題を述べる．

- 今回の実験は小規模であり，大規模データを対象にした場合に有効であるかどうかは明確では

[14] 中心度の算出は特定の議題について行うため，人名をもとに算出した関連度は使用しない．

[15] VLDB だけ低いが，これは，他の 3 会議と比較して，単語の出現頻度の分布が異なるためである．例えば，VLDB で最も出現頻度の高い語は XML であるが，他の 3 会議は web である．

[16] 単語出現頻度を利用した場合，ISWC と LREC が両方とも 1.000 となっているが，さらに細かく見ると，LREC が 1.0000 であるのに対し，ISWC は 0.9998 である．

[17] http://eventseer.net/

[18] http://citeseer.ist.psu.edu/

[19] http://scholar.google.com/

表 3: 各特徴語の出現頻度と出現頻度の高い共起単語対

	ISWC			WWW		
Web	65			152		
ontology	36			7		
RDF	9			3		
OWL	11			5		
XML	2			15		
共起単語対	semantic	Web	55	service	Web	30
	service	Web	26	semantic	Web	25
	ontology	Web	9	search	Web	15
	ontology	semantic	6	application	Web	12
	application	Web	6	content	Web	11

	VLDB			LREC		
Web	7			19		
ontology	0			19		
RDF	1			0		
OWL	0			1		
XML	27			5		
共起単語対	XML	query	9	language	Web	9
	XML	processing	5	semantic	Web	5
	XML	system	3	evaluation	ontology	5
	XML	database	3	resource	Web	4

$$\mathrm{Rel}_{\mathrm{freq}} = \begin{pmatrix} & \text{ISWC} & \text{WWW} & \text{VLDB} & \text{LREC} \\ \text{ISWC} & 1.000 & 0.885 & 0.175 & 0.875 \\ \text{WWW} & 0.885 & 1.000 & 0.323 & 0.735 \\ \text{VLDB} & 0.283 & 0.345 & 1.000 & 0.475 \\ \text{LREC} & 0.944 & 0.741 & 0.267 & 1.000 \end{pmatrix}$$

$$\mathrm{Rel}_{\mathrm{cooc}} = \begin{pmatrix} & \text{ISWC} & \text{WWW} & \text{VLDB} & \text{LREC} \\ \text{ISWC} & 1.000 & 0.397 & 0.000 & 0.292 \\ \text{WWW} & 0.284 & 1.000 & 0.035 & 0.216 \\ \text{VLDB} & 0.000 & 0.238 & 1.000 & 0.000 \\ \text{LREC} & 0.185 & 0.225 & 0.000 & 1.000 \end{pmatrix}$$

図 4: 関連度

表 4: LexRank による中心度

	ISWC	WWW	VLDB	LREC
$\mathrm{Rel}_{\mathrm{freq}}$	1.000	0.956	0.583	0.982
$\mathrm{Rel}_{\mathrm{cooc}}$	0.738	1.000	0.424	0.859
$\mathrm{Sim}_{\mathrm{freq}}$	1.000	0.986	0.715	1.000
$\mathrm{Sim}_{\mathrm{cooc}}$	0.944	1.000	0.836	0.892

ない．今後，国際会議情報を自動取得し，大規模なデータを対象に，その有効性を確認するつもりである．また，大規模なデータを対象にすることにより，特徴語の候補となり得る単語のリストを予め人手で用意せず，IDF 等を利用して特徴語を自動的に選択できるようになると考えられる．

- 小規模な会議の大規模な会議に対する関連度について，その 2 つの会議の間の関連性は低いにも関わらず，高い値が出てしまうことがある．それを防ぐためには，何らかの方法で正規化を行う必要があると考えられる．

- 特徴語を抽出する際，すべての語を公平に扱ったが，例えば,「RDF」という語を題目に含む論文であっても，RDF そのものに関する論文と RDF を使って何か別のことを考える論文では,「RDF」に対する重要性が異なることが考えられる．

- 今回の実験では異なる分野の会議間の関係を分析対象としたが，別の側面として，同一の会議が扱う議題の時間的変化を分析対象とすることも考えられる．

- 第 2 節で述べたように，論文募集要項の内容は，その会議の特徴を示す重要な情報を持っている．しかし，その書式は会議によって様々であり，内容を自動的に分析することは困難であるため，今回の分析対象からは除外した．今後，論文募集要項も分析対象として利用することを考えるべきであると考える[20]．

- 本研究の成果を Web アプリケーションとして公開するためには，会議間の関係を可視化等も必要である．

参考文献

[1] Beate Dorow, Dominic Widdows, Katarina Ling, Jean-Pierre Eckmann, Danilo Sergi, and Elisha Moses. Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination. In *the 2nd Workshop organized by the MEANING Project (MEANING-2005)*, 2005.

[2] Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22, pp. 457–479, 2004.

[3] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *First International Conference on New Methods in Natural Language Processing (NemLap-94)*, pp. 44–49, 1994.

[20] EventSeer は，開催予定の国際会議の論文募集要項を集めたサイトであり，募集要項内の特定の語について，同じ語を含む別の会議と関連付けている．