社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DD－64 (7)
2008－EIP－39 (7)
2008／1／31

ソーシャルメディアへのテキストマイニングの適用に関する検討

大野邦夫 渡辺篤史
職業能力開発総合大学校

本報告は、SNSを代表とするソーシャルメディアから、テキストマイニングを用いて情報を抽出する手法の技術的可能性を検討するものである。テキストマイニングツール、TRUSTIAを用いてmixiのコミュニティの情報から趣味に関する情報を頻度分布として取り出し、それを用いてデータを抽出し、各種コミュニティを相対的に比較した。さらにmixiがサポートしているカテゴリ毎のコミュニティ情報検索機能を用いて、趣味情報の分布を求め、テキストマイニングによる結果との比較を行ない、SNSに対するテキストマイニングの適用領域を考察した。

A Study on Text Mining Application to Social Media

Kunio Ohno, Atsushi Watanabe
Polytechnic University

The goal of this paper is to study the possibility of text mining technology to acquire the information through social media as SNS. A macro program that extracts mixi community information to text mining tool TRUSTIA has been developed.. Vocabulary related to personal hobby of various community have been evaluated through TRUSTIA and statistically compared through histogram. Besides, relationship of hobby vocabulary distribution to community category has been clarified through built−in retrieval function of mixi community, and compared to the text mining result.

1. はじめに

ブログやSNSは、Webの利用者が情報発信を行うことからWeb2.0を代表する分野と考えられ、そのコンテンツの蓄積が進展しているが、その活用法にについては必ずしも十分な検討がなされはいない。ここでは、そのテキスト情報について、テキストマイニングツールを適用する手法を試み、その可能性を検討する。今後のテキストマイニング技術の適用に関しては、昨年の画像電子学会の年次大会でテキストマイニングに関する企画セッションが設けられ、議論されている[1][2][3][4][5][6]。そこでは、研究会報告から研究動向を探る試み[2]や教育分野におけるeラーニング、授業支援などへのアプリケーション[3]が紹介され、さらにWebオントロジを活用する今後の方向性と可能性が論じられた[4][5][6]。

今後の可能性として、個人を特徴付ける情報の抽出が議論された[1]。これは、個人の属性、趣味、嗜好などといった個人情報に基づき、製品やサービスを提供するビジネスを支援する枠組みを構築し、新たなビジネスの可能性の検討を試みるものである。以前、そのような枠組みをPIMオントロジとしてオントロジ技術を使用して推奨するモデルを検討したが[7][8][9]、それを具体化する試みでもある。当然ながら、個人情報保護法を含む社会的なルールにも関係するので、この技術を実用化する際には、そのような問題の解決のための検討も必要になるが、それは別のテーマとして扱いここでは技術的な可能性について検討する。

個人の属性、趣味、嗜好などといった個人情報に基づき、製品やサービスを提供するビジネスとしては、流通分野におけるCRM（Customers Relationship Management）が代表的であるが、放送分野における番組視聴プリファレンスといった分野でも具体的なニーズが顕在化している。さらに教育分野における受講科目選択支援といった分野や、その延長としての個人のキャリアプランへの支援といった分野への適用も考えられる。

以上のような可能性はあるものの、具体的に個人の属性、趣味などにテキストマイニングを適用した事例はあまり多いとは言えない。そもそもテキストマイニングを個人を特徴付ける情報抽出に適用した事例やノーハウが乏しいように思われる。そこで、個人を特徴付ける意味で扱いやすいと思われる個人の趣味情報について、mixi[10]を用いて検討することとした。

2. mixi活用の可能性

2.1 mixiの基本機能

以上述べたような背景に基づき、代表的なSNSであるmixiを対象に、テキストマイニングの活用を試みる

こととした。テキストマイニングツールとしては、ジャストシステムのTRUSTIA[11]を用い、その機能の範囲で可能なことを検討する。

mixiには、2007年11月末の時点で1290万余りの登録された参加者があり、趣味を含む個人プロファイルが表示されているだけでなく、そこで日記を書いたり、書籍、CD、DVDなどのレビューを推薦する機能を持つ。さらに参加者が自由に作成するコミュニティが存在し、そこでは掲示板にトピックを作成したり、そのコミュニティに関するイベントを企画し参加者を募ることが可能である。参加者のアカウントには、参加コミュニティが表示されるが、そのアカウントの参加者が関心を持つコミュニティを通じてその人の関心領域を知ることが可能である。さらにマイミクと呼ばれる友人のリストが表示され、交友関係を知ることができる。以上のとおり個人に着目すると、これらの情報だけでその人の人となりをかなり知ることができる。

2.2 コミュニティ情報の活用

コミュニティに着目すると、別の観点から興味深い情報を知ることができる。mixiのコミュニティは、表1のように分類され、各々のカテゴリにおいてさまざまなコミュニティが存在する。

テキストマイニング技術が期待される適用領域としては、日記、レビュー、コミュニティの掲示板が考えられる。それらの文章から特徴的なキーワードを抽出し、その日記の著者レビュー対象やコミュニティ掲示板、さらにそれらへのコメントに関する特徴を抽出することが可能である。

2.3 テキストマイニングの適用領域

日記の内容についてテキストマイニングを適用すると、その著者が好むキーワード群が系統的に抽出される。それらのキーワードを自己紹介で記述されたプロファイル情報と対応付けたり、参加コミュニティと対応付けることにより、その個人についてより具体的な情報を得ることが可能である。しかしそのような個人情報は、その個人の了解なしに公開することはできないので、研究対象とするには制約がある。本人の了解の下に匿名として統計的に扱うような場合にのみ可能であろう。

コミュニティ掲示板やレビューは、日記に比べると制約は少ないと思われる。かつコミュニティ掲示板はレビューに比べると分類体系が整っている。そこでコミュニティ掲示板のトピックとコメントについてテキストマイニングを行うことにより、そのコミュニティを特徴付ける意味分類のキーワードを抽出することを試みた。

表1　mixiのコミュニティ分類

分類	コミュニティカテゴリ
娯楽	音楽、映画、スポーツ、ゲーム、本・マンガ、旅行、車・バイク、占い、趣味、動物・ペット
知識	PC・インターネット、学問・研究、ビジネス・経済、アート
生活	地域、グルメ・お酒、ファッション
グループ	学校、会社・団体、サークル・ゼミ、同年代
芸能	芸能人・有名人、テレビ番組、お笑い
その他	その他、イベント

2.4 TRUSTIAのカスタマイズ

mixiのコミュニティ掲示板の情報を、TRUSTIAで分析するために、テキスト抽出のインタフェースプログラムを開発した。TRUSTIAは、図1に示すように教師支援するために教育用にカスタマイズされたツールで、マイニング機能を持つマイニングアシステントとその支援環境であるティーチングアシスタントから構成されている。TRUSTIAをmixiに適用するためには、図1におけるティーチングアシスタント環境を、

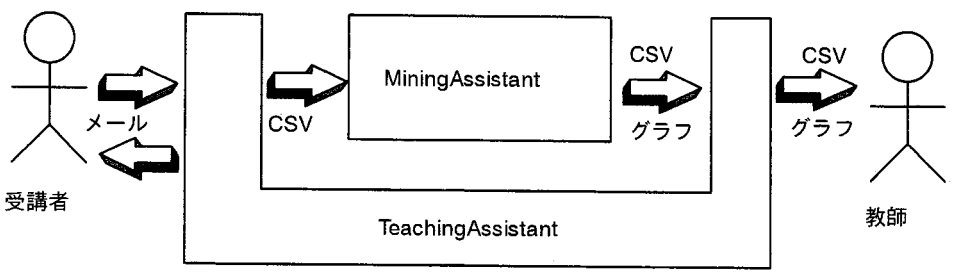

図1　TRUSTIAのシステム構成

mixiからデータを抽出するための環境に適合させることが必要である。この環境は、CSV形式のデータをサポートしており、ExcelとVBAによるマクロでカスタマ

イズ可能である。この機能を用いて、mixiの掲示板内容を取り込むことを可能とした。

TRUSTIAを使用するには、図2に示すように、データ収集 → データ加工 → 辞書登録 → 統計解析 → 主題分析 → 傾向把握 → 表現抽出 → 属性把握 → 語彙分析といったテキストマイニング・プロセスを経る。

3. TRUSTIAによる検討

3.1 分析の方法

上記手法の適用例として、mixiのコミュニティにおける個人の趣味の分析を試みた。mixiのプロファイルにおいて趣味に関する分類で用いられている用語群を抽出して、コミュニティ参加者が持つと思われる趣味の方向性を分析した。その手順は下記のとおりである。

(1) すべてのカテゴリから、人数が多いコミュニティの順に3つのコミュニティを選択する。

(2) 作成したプログラムを用いそのコミュニティの掲示板のトピック及びコメントからmixiの趣味分類毎にテキストを抽出する。

(3) TRUSTIAを用いて、テキストマイニングを行い、抽出されたキーワードを含む文章解析を行う。

(4) 抽出されたキーワードの意味の出現回数をコミュニティ毎にリストアップする。

将来的にはコミュニティ参加者の個々人とコミュニティの性格を関係づけることも検討したいが、時間的な制約で今回はそこまでの分析は差し控えることとした。

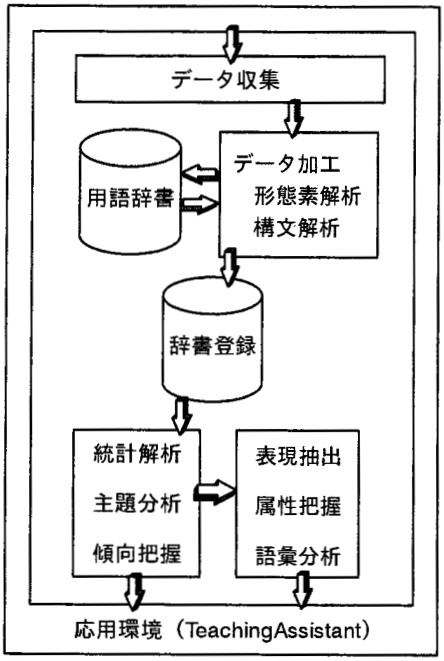

図2　MiningAssistantの構成

3.2 TRUSTIAによる分析結果

前項の手順で得られたマイニング結果の例を、表2に示す。表1の大分類における最初のコミュニティについての結果を示している。

表2　コミュニティから抽出されたキーワード順位（その1）

カテゴリ	コミュニティ	趣味－1	趣味－2	趣味－3
音楽	Mr.Children	カラオケ・バンド	ドライブ	テレビ
	音楽がないと生きて	カラオケ・バンド	音楽鑑賞	ドライブ
	おすすめ洋楽	カラオケ・バンド	語学	インターネット
PC・インターネット	かわいい顔文字	インターネット	アウトドア	ペット
	ショートカットキー	インターネット	読書	漫画
	You Tube	インターネット	テレビ	ゲーム
地域	関西人	ファッション	旅行	カラオケ・バンド
	大阪この店あの店	ショッピング	お酒	ファッション
	I love yokohama	ドライブ	ショッピング	ファッション
学校	授業中に寝てました	習い事	語学	インターネット
	試験をなめている	漫画	語学	ゲーム
	ミクナビ◆［就職活動］	インターネット	テレビ	語学
芸能人・有名人	安室奈美恵のPV	カラオケ・バンド	ファッション	テレビ
	浜崎あゆみ	カラオケ・バンド	ファッション	ドライブ
	mixi登録有名人	インターネット	テレビ	カラオケ・バンド

この結果を見ると、カラオケ・バンド、インターネット、ファッション、テレビといったキーワードが多いが、これらのキーワードが全般的に多いわけではない。キーワードには、コミュニティ特有のものと、

コミュニティ横断的なものとがあり、それらを分析することにより、個々のコミュニティの性格や動向を知るヒントになる。

スペースの都合で各コミュニティについてキーワードを3個までしか記載していないが、今回の検討では5個まで採取した。それらは各コミュニティにおける出現回数として得られている。因みにそれらを出現回数順に列挙すると、ゲーム（3147）ドライブ（3056）、ペット（2577）、読書（2092）、アート（1854）、スポーツ（1557）、テレビ（1547）、漫画（1376）、美容・ダイエット（1311）、旅行（1153）、料理（1134）、インターネット（779）、カラオケ・バンド（715）、ショッピング（435）、ファッション（366）、スポーツ観戦（349）、語学（266）、お酒（211）、映画鑑賞（153）、習い事（96）、音楽鑑賞（91）、グルメ（56）、ギャンブル（54）、アウトドア（8）であった。

表3　コミュニティから抽出されたキーワード順位（その2）

カテゴリ	コミュニティ	趣味－1	趣味－2	趣味－3
映画	映画愛好会	読書	映画鑑賞	テレビ
	Jack Sparrow	テレビ	漫画	カラオケ・バンド
	おすすめ映画	映画鑑賞	読書	テレビ
学問・研究	キュンとする言葉	漫画	テレビ	アート
	ことばのくすり	読書	テレビ	漫画
	mixiで使える顔文字	インターネット	語学	漫画
グルメ・お酒	関素材別★簡単おいしい	料理	お酒	読書
	ラーメン大好き	カラオケ・バンド	旅行	読書
	I love yokohama	ドライブ	ショッピング	ファッション
会社・団体	Tokyo Disney Resort	旅行	ドライブ	テレビ
	よく物をなくす	ドライブ	読書	テレビ
	妄想族！	テレビ	ドライブ	読書
テレビ番組	恋愛観察バラエティ	テレビ	読書	旅行
	ピタゴラススイッチ	テレビ	スポーツ	読書
	水曜どうでしょう	テレビ	ドライブ	旅行

表3は、表1の大分類における2番目のコミュニティについての結果を示している。表2と比べると分かるが、キーワードはかなり異なっている。このことから、異なっているキーワードは、コミュニティへの依存度が大きいのに対し、どのコミュニティにも比較的頻繁に出現するキーワードは一般的に関心を持たれる趣味であると言える。

3.3 キーワードのコミュニティ依存性

キーワードのコミュニティ依存性を評価する方法を考えてみた。今回の検討では、72のコミュニティを対象にしている。データはExcelの表で管理しているが、個々のキーワード毎にソートしてみた。その結果を表4に示す。

表の列は、出現回数の多い順に並んでいるが、分布形状はかなり異なることが分かる。ゲームやドライブをはじめとする多くのキーワードは、ピークのコミュニティで高い頻度で出現し、他のコミュニティについてはその頻度は急速に減少するが、テレビやインターネットに関しては、ピークのコミュニティに比較して減少の程度がなだらかである。このような、ソーティングによるキーワードのコミュニティ依存性を調べることにより、そのキーワードの種々の分野における認知性や受容性が明らかになると考えられる。

以上の通りキーワードの出現頻度はコミュニティに強く依存しているが、TRUSTIAにおいては、それらのキーワードに関係する語彙群による主題が、構文解析された文章相互の関係を通じて木構造で管理されている。より詳細な分析を行う場合には、このような情報も活用可能と思われる。

4. mixiのコミュニティ検索機能の活用

4.1 分析方法

mixiのコミュニティのキーワード検索機能を用いると、図3で示すように各々のカテゴリにおいてヒットするコミュニティ数が表示され、さらにヒットしたコミュニティが、作成日時順又はメンバー数順に表示される。この図では、学校関係のカテゴリのコミュニティ群において紹介文にカラオケという文字列を含むコミュニティが668存在することを示している。このような簡単な操作で各種カテゴリにおいて特定のキーワードを含むコミュニティ数を知ることが可能である。カテゴリは、表1で紹介した通りであるが、メニューから学校、会社・団体、サークル・ゼミ、地域、同年代、音楽、スポーツ、ゲーム、映画、本・漫画、アート、テレビ番組、芸能人・有名人、お笑い、PC・インターネット、グルメ・お酒、旅行、動物・

表4　コミュニティ毎のキーワードの頻度分布

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ゲーム	ドライブ	ペット	読書	アート	スポーツ	テレビ	漫画	美容・	旅行	料理	インタ	カラオ	ショッピ	ファッ	スポー	語学	お酒
2	2347	1459	1024	1367	1609	966	314	488	946	726	974	180	204	79	56	163	71	64
3	608	1166	786	165	82	175	253	217	156	214	55	62	90	45	52	114	33	19
4	44	43	634	71	14	89	161	185	75	18	18	56	82	29	40	61	29	14
5	23	35	10	40	14	60	77	170	26	17	16	42	68	28	29	2	23	9
6	21	34	7	34	11	48	73	51	14	16	13	35	43	25	23	2	19	8
7	11	24	7	30	11	22	72	37	6	16	8	26	41	24	18	2	16	6
8	9	21	7	27	9	18	46	32	6	15	5	26	15	13	17	1	7	5
9	9	21	6	26	8	16	37	22	5	14	5	20	12	13	10	1	6	5
10	7	21	6	23	8	13	35	20	5	9	4	19	11	12	9	1	4	5
11	7	18	5	21	7	12	34	12	4	9	3	19	11	11	8	1	3	5
12	5	18	5	16	7	11	32	10	4	8	3	14	10	11	8	1	3	4
13	4	18	5	15	7	10	31	8	4	8	2	13	10	10	7	0	3	4
14	4	17	5	13	6	10	26	7	4	6	2	12	10	10	6	0	3	3
15	4	11	5	13	5	8	26	6	4	6	2	11	10	9	6	0	3	3
16	3	11	4	12	5	7	25	6	4	5	2	11	8	7	5	0	2	3
17	3	10	4	12	4	6	24	6	4	5	2	11	6	7	4	0	2	3
18	3	10	4	11	3	5	21	6	3	5	2	10	6	7	4	0	2	3
19	2	9	4	11	3	5	19	5	3	4	1	10	5	6	4	0	2	3
20	2	8	4	10	3	4	17	5	3	4	1	10	5	6	4	0	2	3

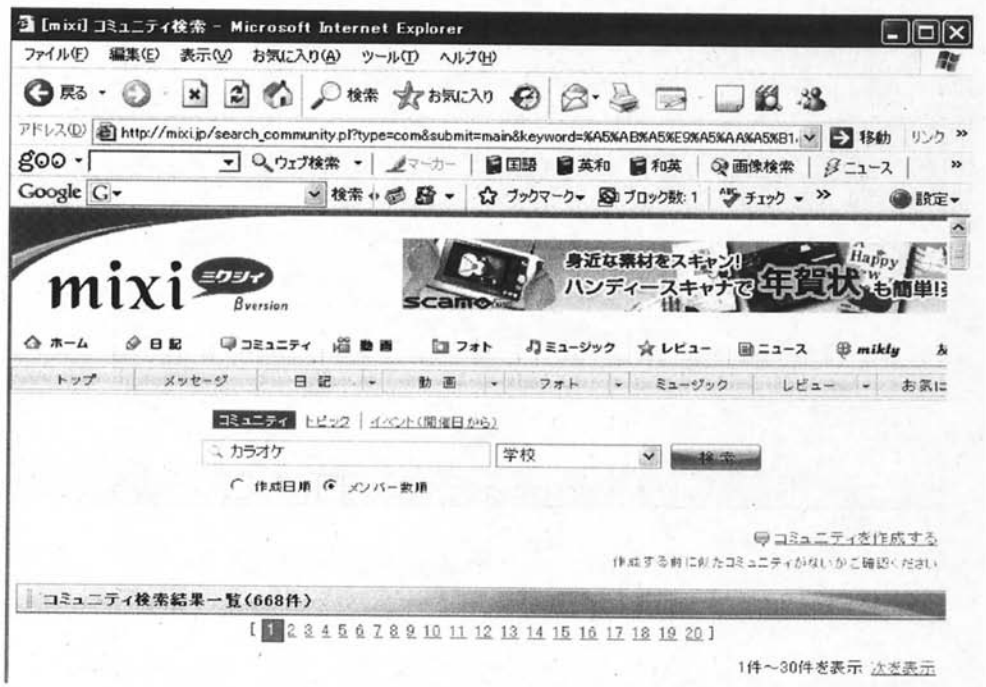

図3　mixiにおけるコミュニティ検索画面

ペット、車・バイク、ファッション、占い、趣味、ビジネス・経済、学問・研究、アダルト、その他といった28種類の硬軟取り混ぜた多様なバラエティを選択可能となっている。作成される、各々のコミュニティはその何れかのに属している。

これらのコミュニティの紹介で用いられているテキストの語彙を検索し、比較することにより、各コミュニティの相対的な性格を知ることが可能である。さらに、種々の語彙について検索すると、コミュニティ毎の語彙に対する分布が抽出される。この語彙に、前項

で検討した個人プロファイルにおける趣味の分類を適用すると、趣味に関するキーワードとコミュニティの関係を知ることができる。

4.2 カテゴリ毎の検索結果

前項のようにして検索した結果をExcel上にまとめたデータを表5に示す。この結果を用い、表2、表3で紹介したカテゴリにおいてヒットしたキーワードの順位を表6に示す。

表5　mixiの検索機能によりヒットしたカテゴリ毎のコミュニティ数

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	映画	鑑スポー	スポー	音楽	鍵カラオ	料理	グルメ	お酒	ショッピ	ファッシ	アウトド	ライブ	旅行	アート	習い事	語学	読書
学校	18	1189	6	6	1462	393	95	361	78	518	61	131	1153	435	44	1071	103
会社、団体	28	1090	10	9	1882	842	239	697	233	466	179	212	1311	549	35	225	77
サークル、	196	996	136	51	2080	1055	1128	1057	314	1176	1250	1238	1232	854	281	847	389
地域	52	844	43	11	1490	874	814	902	887	977	524	893	808	690	266	404	101
同年代	14	398	11	8	905	256	115	443	67	337	83	148	1165	68	58	68	34
音楽	132	585	18	132	1524	835	125	956	175	820	109	922	649	835	31	83	186
スポーツ	2	1498	28	35	19	8	5	557	4	12	7	8	18	7	0	2	3
ゲーム	0	246	1	0	217	70	0	23	13	3	0	165	6	6	0	13	46
映画	1	66	0	0	253	342	21	46	33	12	9	93	159	303	2	29	35
本、漫画	10	10	0	3	440	224	80	76	33	296	24	124	305	242	6	51	672
アート	0	9	0	4	785	380	56	198	44	1085	34	80	419	1350	27	30	76
テレビ番組	9	18	0	4	334	400	102	34	61	13	10	67	195	81	2	44	38
芸能人、有	305	67	0	35	2396	1134	173	320	355	1376	42	325	586	445	21	191	586
お笑い	27	241	0	50	736	277	72	2	40	174	26	125	349	103	5	51	87
PC,インター	4	6	1	5	224	80	56	3	125	103	17	114	150	154	0	38	30
グルメ、お	3	573	3	5	1890	1292	1376	1388	230	258	266	358	1086	448	23	44	98
旅行	0	41	2	0	140	722	399	176	177	78	220	429	1331	113	5	226	32
動物、ペット	0	2	0	4	141	61	18	64	28	111	34	48	224	51	4	7	14
車、バイク	0	1367	3	6	191	68	6	112	5	167	136	1456	262	128	1	5	19
ファッション	2	44	6	7	661	201	116	154	703	1131	194	78	326	1385	29	14	47
占い	2	25	0	4	59	29	9	1	3	29	4	5	57	35	2	3	5
趣味	46	39	14	30	2250	1438	428	1008	255	928	643	600	1454	974	141	177	269
ビジネス、	4	152	3	3	305	220	2	108	206	273	24	43	462	159	27	100	36
学問、研究	0	281	6	9	631	531	140	319	59	347	55	141	724	449	27	832	285
アダルト	2	2	7	12	405	141	54	6	48	150	31	69	210	119	1	9	32
その他	1	1025	33	34	2691	84	542	71	395	1409	326	586	1396	1038	99	245	285
総計	858	10814	331	467	24111	11957	6171	9082	4571	12249	4308	8458	16037	11021	1137	4809	3585

表6　カテゴリ毎のキーワードの順位

カテゴリ	趣味−1	趣味2	趣味−3	趣味−4
音楽	カラオケ・バンド	お酒	ドライブ	ペット
PC・インターネット	インターネット	ゲーム	テレビ	カラオケ・バンド
地域	カラオケ・バンド	ゲーム	美容・ダイエット	ファッション
学校	カラオケ・バンド	スポーツ	旅行	語学
芸能人・有名人	カラオケ・バンド	ゲーム	ファッション	テレビ
映画	テレビ	ゲーム	料理	ファッション
学問・研究	ゲーム	語学	テレビ	旅行
グルメ・お酒	カラオケ・バンド	お酒	グルメ	料理
会社・団体	カラオケ・バンド	旅行	ゲーム	スポーツ
テレビ番組	テレビ	ゲーム	漫画	料理

表2、表3で出現したキーワードと多少類似している面もあるが、かなり異なると言える。その理由は、表2、表3が、そのカテゴリにおける特定の（人数が多い順で3つの）コミュニティを対象にしたのに対し、mixiのコミュニティ検索機能は、カテゴリにおけるすべてのコミュニティを対象とするためである。さらに、TRUSTIAのマイニング機能を用いた表2、表3の結果が、コミュニティの紹介文だけでなく、トピックの内容やコメントを含むテキストを対象としているのに対し、表6ではコミュニティの紹介文のみを対象とするためであろう。さらに、表2、表3では、TRUSTIAによる同義語、類義語などの処理が付加されているのに対し、表6では文字列一致しか対象としていないことも一つの要因として挙げられるであろう。

4.3 カテゴリ毎のキーワードの分布

先に表4で、キーワードのコミュニティ依存性について示したので、同様の手法でカテゴリ依存性について

データを出してみた。Excel上にまとめたデータを表7に示す。

表7　キーワードのカテゴリ依存性

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ペット	読書	アート	スポー	テレビ	漫画	美容	旅行	料理	インタ	カラオ	ショッ	ファッ	スポー	語学	お酒
2	1384	672	1385	1498	1230	1436	1237	1454	1438	1447	2396	887	1376	136	1071	1388
3	892	586	1350	1367	1211	1242	1144	1331	1292	1183	2250	703	1176	43	847	1057
4	788	389	974	1189	1165	1205	925	1311	1134	858	2080	355	1131	28	832	1008
5	739	285	854	1090	1096	840	922	1232	1055	665	1890	314	1085	18	404	956
6	536	269	835	996	1089	607	782	1165	874	531	1882	255	977	14	226	902
7	525	186	690	844	1021	570	727	1153	842	500	1524	233	928	11	225	697
8	483	103	549	585	872	492	504	1086	835	459	1490	230	820	10	191	557
9	351	101	449	573	839	393	414	808	722	432	1462	206	518	6	177	443
10	337	98	448	398	814	390	396	724	531	403	905	177	466	6	100	361
11	288	87	445	281	781	360	327	649	400	274	785	175	347	6	83	320
12	277	77	435	246	753	329	291	586	393	237	736	125	337	3	68	319
13	271	76	303	241	704	279	233	462	380	235	661	78	296	3	51	198
14	261	47	242	152	549	275	223	419	342	208	631	67	273	3	51	176
15	213	46	159	67	464	247	153	349	277	187	440	61	258	2	44	154
16	155	38	154	66	368	246	135	326	256	172	334	59	174	1	44	112
17	155	36	128	44	350	199	76	305	224	169	305	44	167	1	38	108
18	152	35	113	41	348	150	67	262	220	153	253	40	111	0	30	76
19	149	34	103	39	297	138	56	224	201	150	224	33	103	0	29	64
20	116	32	81	25	212	96	53	195	80	85	217	33	78	0	14	46

表4におけるコミュニティでの頻度分布に比べると、カテゴリ毎のコミュニティ数は幅広く分布している。それでもペット、テレビ、旅行、カラオケといったキーワードは、特に分布がゆるやかであり、カテゴリ横断的に用いられる趣味であることが分かる。

5. 考察

5.1 結果の要約

今回は手始めにmixiのコミュニティ掲示板に、テキストマイニングツールのTRUSTIAを適用し、その感触と可能性の把握を試みた次第である。十分な検討とは言い難いが、下記の結果が得られたと考える。

(1) mixiのコミュニティの掲示板からトピックとコメントを構造を持ったテキストとして抽出し、個別の構造毎に形態素解析・構文解析を行い、主題分析・語彙分析に基づきキーワードの出現頻度を得た。

(2) キーワードの出現頻度はコミュニティに強く依存し、そのキーワードに関係する語彙群が木構造表現として得られた。

(3) mixiのコミュニティ検索機能をを用い、カテゴリ毎にキーワードを含むコミュニティ数を算出した。この情報はキーワードのカテゴリ依存性を把握するためには有効である。

(4) キーワードのカテゴリ依存性は、コミュニティ依存性に比べると比較的ゆるやかであるが、それでもキーワードによる依存性には差異があるので一つの指標とすることが可能である。

5.2 今回の到達点

mixiの日記、レビューとコミュニティ掲示板を対象に、テキストマイニングを試みることを出発点にして、個人情報を扱わねばならない日記や分類体系があいまいなレビューについては後回しにして、今回はコミュニティ掲示板のみを対象にした。そのため、かなり限られた検討になってしまったが、mixiが本来持っているコミュニティのカテゴリを活用する手法との比較を行うことにより、テキストマイニングの役割が明確になったと考えられる。

すなわち、mixiという巨大な人口を擁するコミュニティがあり、そこに含まれる情報を活用するためには、その全体を眺める巨視的な視点から、個々人を対象にする微視的な視点まで、幅広い視点を必要とする。巨視的な視点にとっては、コミュニティのカテゴリが便利な指標であり、微視的な視点では、個々人のプロファイルや日記が有効である。

個々のコミュニティは、その中間的な存在であり、巨視的な視点と微視的な視点が交差する非常にアクティブな場であると言える。今回の検討で不十分ながらテキストマイニング技術がこの分野に適合しそうなことが判明したと言えるであろう。

なお、巨視的な視点を与えるmixiのコミュニティ検索機能は、簡単に使える非常に便利な動向分析ツールである。テキストマイニング以前に、このコミュニティ検索機能を用いてマクロな視点における知見を得た後に、テキストマイニングで詳細な分析を行い、さらにそれを個々人のデータで確認するといった手法がリーズナブルな方法論として考えられるのではないかと思われる。

5.3 反省点

今回は、個人プロファイルとコミュニティを結びつける情報を取り上げるということで、趣味関連の語彙をキーワードとする分析を行ったが、方法論的な検討しかできず、意味内容までは議論できなかった。その最大の理由は、mixiの扱う分野がきわめて膨大であり、テキストマイニングの対象領域を絞れなかった点にある。趣味といえども、そのカテゴリは膨大であり、テキストマイニングの具体的な目的を設定しカテゴリを絞らなければ有効な結論を得られないことを痛感した。

要するに、一般的な個人の趣味から出発するようなボトムアップ的なアプローチでは適用分野を絞れなかったと言える。長期的には、CRMやTV視聴ブリファレンスサービスを視野に置いたPIM関連情報の活用を目標としているが、そこに至るためには、目標を逐次明確化して行くトップダウンによる具体化のステップを考える必要がある。

具体的な目標は、当面の実用に役立つ技術的なシーズや市場ニーズのようなものが望ましいと考えられ、mixiのコミュニティで活発に議論されているホットな話題について、分析できるようになれば興味深いと考えられる。

6. 今後の課題

以上に基づき、今後は以下のような検討を進めたいと考える。

6.1 適用領域の検討

長期的には、CRMやTV視聴ブリファレンスサービスを視野に置いたPIM関連情報の活用を目標とするが、そのステップとして、具体的な対象を絞り込みたい。取りあえずは、当研究室で検討している情報家電向けのコンテンツメタデータやIPネットワークメタデータに関する分野において、パーソナライズ機能の活用可能性について、ユースケース・利用シナリオを検討し、そのモデルに適合するデータ取得の検討を行いたいと考える。

6.2 mixiのコミュニティ検索機能の活用

テキストマイニングを行う以前に、大まかな動向把握のためのmixiにビルトインされたコミュニティ検索機能は便利でありこれを有効に活用することを検討したい。この情報に基づき、実用的な目標を定め、その実現のためのテキストマイニングの具体的なアプローチを設定する。

6.3 テキスト抽出プログラムの機能拡張

今回は、mixiのコミュニティの掲示板のトピックとコメントを抽出するプログラムを作成したが、これを日記やレビューにも適用できるように機能追加を行う。

7. 終わりに

以上、日本の代表的なSNSであるmixiのコミュニティに、テキストマイニングを適用して検討し、その可能性と課題について述べた。技術的な目処が立っていたので、有効な結果が容易に得られるかと考えたが、適用に当たってはより具体的な目標設定が必要であることが判明した。そのためには目標を設定した上でのトップダウン的なアプローチが必要であるが、そのためにmixiのコミュニティ検索機能を活用することが可能であることも判明した。

最後に、本研究は、職業能力開発総合大学校通信システム工学科における卒業研究としてなされたものであることを述べるとともに、テキストマイニングツールTRUSTIAを無償で貸与して下さった(株)ジャストシステムに謝意を表します。

文献

[1] 大野; "テキスト検索から意味的情報マイニングの時代への展望", 2007年度画像電子学会第35回年次大会講演論文集【T.4−1】(2007.6)

[2] 斎藤; "テキストマイニングによる研究技術動向の検討", 2007年度画像電子学会第35回年次大会講演論文集【T.4−2】(2007.6)

[3] 石沢; "テキストマイニングによる学習支援", 2007年度画像電子学会第35回年次大会講演論文集【T.4−3】(2007.6)

[4] 細見; "程度表現オントロジの提案：Webマイニングのためのレイティング用語彙の考察", 2007年度画像電子学会第35回年次大会講演論文集【T.4−4】(2007.6)

[5] 西岡, 濱田, 鬼塚, 山室; "XMLを用いた多者間取引支援システム − データベース検索とストリーム処理の融合マッチング− ", 2007年度画像電子学会第35回年次大会講演論文集【T.4−5】(2007.6)

[6] 新, 木谷, 木村, 新, 大野; "ネットワークの操作・管理におけるXMLメタデータ応用", 2007年度画像電子学会第35回年次大会講演論文集【T.4−6】(2007.6)

[7] 大野; " オントロジ技術の応用に関する一考察", 情報処理学会デジタルドキュメント研究会講演論文, DD41.1, (2003.9)

[8] 大野; " コンパウンドドキュメントにおけるモバイル・パーソナルプロファイル", 画像電子学会VMA研究会講演論文, (2006.1)

[9] 大野; " モバイルCRMへのオントロジ適用の可能性", 画像電子学会VMA研究会講演論文, (2004.1)

[10] http://mixi.jp/

[11] 大野, 石沢, 横田; "パーソナル・テキストマイニング技術の可能性", 画像電子学会VMA研究会講演論文, (2007.1)