3E2-09

クラスタリングに基づく概念表現と TREC Novelty Track における評価

Conceptual expression based on clustering, and evaluation in TREC Novelty Track.

角田 祐一　　新村 昭好*　　高木 友博
Yuichi Kakuta　　Akiyoshi Sinmura　　Tomohiro Takagi
明治大学 理工学研究科 基礎理工学専攻
Department of Computer Science, Meiji University

Abstract: We propose conceptual fuzzy sets based modeling to express the meaning of sentences and words which vary depending on contexts. We examined three approaches to build conceptual fuzzy sets, clustering, SVD and WordNet, applying them to TREC 2003 Novelty Track corpus. The clustering approach resulted in 5th in the track to distinguish relevant sentences from non-relevant ones.

1. はじめに

情報検索などの研究分野では，検索の精度を向上させるために，クエリや文書の持つキーワードを拡張させることが多い．しかし，キーワード拡張にはシソーラスのような語の上位・下位関係や類似関係，連想関係などが整理されているものを用いなくては、精度の向上は望めない．ところが，現実には，語の上位・下位関係は，正確に定義することが困難であることに加え，状況に応じてその意味を変化させる．

本研究では，状況に応じた意味を捉えるために，語・文の曖昧な意味(概念)を他の言葉によって表現する概念ファジィ集合(CFSs: Conceptual Fuzzy Sets)を用い、状況に依存した概念拡張のモデル化を試みる．

2. 概念ファジィ集合

概念ファジィ集合とは，言葉の意味の状況依存性を志向したファジィ集合で，概念の意味するファジィ集合を関連する言葉の活性度の分布により表現すると同時に操作を内包する知識表現である．まず，核になる概念とそこから連想される言葉を連想記憶上にあらかじめ記述しておく(これを辞書と呼ぶこととする)．ここである概念に対応する言葉が活性化されると，活性値伝播が起こり，その概念の意味が想起される．この活性値伝播により，状況に依存して変化する概念の意味表現が可能になる．

例として図 1 の状況を考える．"Java"には「コーヒー」「島」「プログラム言語」と複数の意味がある．しかし，同じ文書内に"Hard disk"が出現していれば，文脈は"Computer"であると認識することができる．この"Computer"が核になる概念である．次にこの概念から連想できる言葉に活性値を伝播することにより，"Programming language"や"Pemtium4"などの単語が活性化され，これらの言葉の活性値分布でこの文脈の意味を表現する．

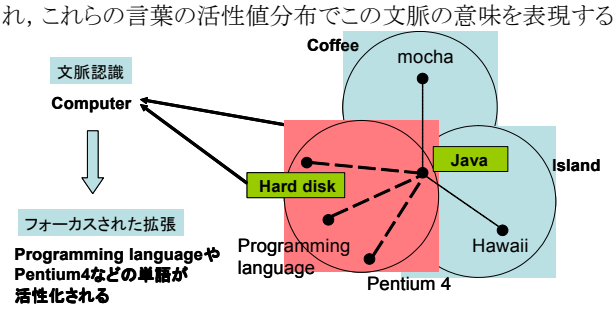

図1　概念ファジィ集合による想起

3. 状況依存動的概念拡張モデル

状況依存的概念拡張を概念ファジィ集合によって実現する．概念拡張モデルとしてニューラルネットワークの一種である RBF (Radial Basis Function) ネットワークを参考にした．概念拡張モデルを図 2 に示す．

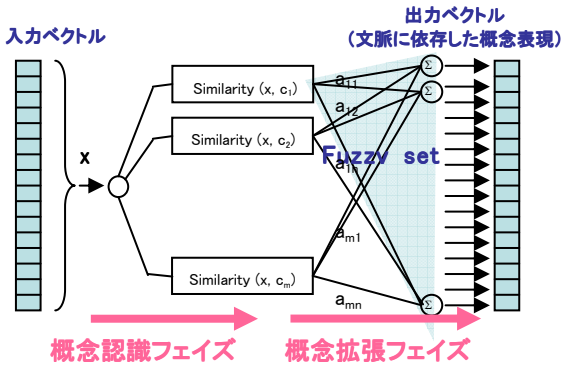

図 2　概念拡張モデル

概念拡張は 2 段階の過程で行う．まず，核となる概念(中心概念)c_i と入力 x との類似度を計算する．類似度を中心概念 c_i の活性度とする．この過程を，入力 x に近い概念が活性化するという点で概念認識フェイズとする．次に，中心概念 c_i の活性値を関連する単語 j に重み a_{ij} を掛けて伝播する．この過程を，認識した概念の意味を複合処理するという点で概念拡張フェイズとする．入力 x に対する概念拡張後の単語 j の活性値 $f_j(x)$は以下の式(1)によって計算される．

$$f_j(x) = \sum_i^m a_{ij} Similarity(x, c_i) \qquad (1)$$

4. 概念辞書構築

上記の概念拡張を行うためには，知識となる辞書が必要である．提案する概念拡張モデルにおける辞書は中心概念 c_i と重み a_{ij} に相当する．

4.1 中心概念 Ci の構築

本研究では，3 種類の中心概念の構築手法を提案し比較，検証する．

連絡先：角田　祐一，明治大学 理工学研究科 基礎理工学専攻，214-0034 川崎市多摩区東三田 1-1-1，044-934-7483，kakuta@cs.meiji.ac.jp

*現在，沖電気工業株式会社に所属

(1) クラスタリングによる構築

文書コーパスを K-Means アルゴリズムによりクラスタリングし，そのクラスタの重心ベクトルを概念とした．クラスタはある話題についての文書集合であり，その重心ベクトルはその話題についての概念を表現しているものである．本研究ではクラスタ(中心概念)数を経験的に 800 と 1600 とした．

(2) 主成分分析による構築

文書コーパスから単語の文内共起を求め，その情報を単語間の関連度，類似度とみなし，主成分の分析を行った．

主成分の分析は，SVDPACKC**を用いて行った．その結果得られた上位の主成分を中心概念として設定した．

(3) WordNet による構築

WordNet は人手で作成されているため，上位概念，下位概念など，語(概念)と語(概念)の関係がきちんと整理されている．WordNet の構成要素は同義語集合で形成された概念である．ここではこの概念の中から約 1 万個の概念を選択し，中心概念として設定した．

4.2 類似度計算

(1) c_i がクラスタリングによる概念の場合

中心概念 c_i と入力 x とのコサイン尺度

$$Similarity(x, c_i) = \frac{x \cdot c_i}{\| x \| \| c_i \|} \qquad (2)$$

を用いる．

(2) c_i が主成分分析による概念の場合

(1)と同様に，式(2)を用いる．

(3) c_i が WordNet による概念の場合

中心概念がベクトルの形式をしていないため，式(1)を用いることができない．WordNet の特徴である上位・下位などの定義された関係を利用し，入力単語の活性値を上位概念へ伝播，下位概念へ伝播させ，さらに排他的拡張(負の活性値伝播)を行う．

4.3 中心概念 c_i-出力層の単語 j 間の重み a_{ij} の学習

概念拡張フェイズで利用する中心概念とそれに関連する単語間の重みは，最小二乗法により学習を行った．

学習に必要となる教師データは,文書コーパスに用意されているトピックと文書を用いて作成した．

WordNet による中心概念の構築の場合，4. 2 で述べたようにベクトル化されていないため,重みを定義することができない．よって,中心概念の類似度計算での操作でをそのまま出力とした．

5. 検証システム

提案した概念拡張手法を検証するために情報フィルタリングシステムを構築し，TREC(Text REtrieval Conference)で行われている*†タスクを実行させた．各概念拡張手法をこのタスクに適用した精度で比較する．

行ったタスクは，あるトピックが与えられ，それに適合する文を文書ストリームから見つけるという情報フィルタリングのタスクである．構築したフィルタリングタシステムの概要を図 3 に示す．

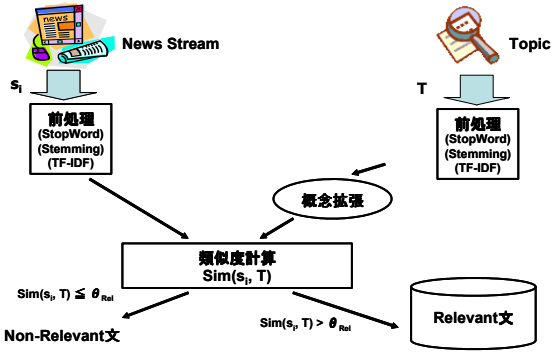

図 3 フィルタリングシステム

まず，与えられたトピック T からプロファイルを作成する．一般的な前処理によりトピックをベクトル化し，そのベクトルに対して概念拡張を行ったものをプロファイルとする．また，入力文 s_i は前処理によりベクトル化し，プロファイルとの類似度を計算する．類似度が閾値以上であればトピックに適合する文とし，閾値以下であれば適合しない文とする．

6. 検証結果・考察

6.1 TREC タスク実行結果

3 つの手法で作成した中心概念で構成したシステムを TREC のタスクに適用した結果を表 1 に示す．

表 1 概念拡張手法別の結果

概念拡張アプローチ	Run	Recall	Precision	F-Measure	Improvement (vsMeijiHilF15)	備考
tf-idf	MeijiHilF15	0. 55	0. 57	0. 498	-	
共起	un-official	0. 62	0. 56	0. 53	6. 85%	
WordNet	un-official	0. 31	0. 6	0. 363	-26. 82%	上位概念への拡張
	un-official	0. 31	0. 6	0. 362	-27. 02%	上位->下位概念への拡張
	un-official	0. 32	0. 6	0. 367	-26. 01%	上位->下位概念->排他的拡張
主成分分析	un-official	0. 43	0. 5	0,411	-17. 47%	概念数 1213
クラスタリング	MeijiHilF13	0. 84	0. 52	0. 589	18. 70%	概念数 800
	un-official	0. 57	0. 47	0. 46	-7. 64%	概念数 1600

表 1 において Run 列は TREC で結果を識別するタグである．MeijiHilF と付く結果は TREC に実際に提出した結果であり，un-official と付く結果は，その後検証として行った結果である．

表 1 の結果を見るとクラスタリングアプローチによる概念拡張がもっとも高い F-Measure を示している．

概念拡張の検証が目的であるため，un-official と付く結果もシステムのそれ以外の部分は全て TREC に提出したとき同じでものである．MeijiHilF13 の結果は 検証に用いたタスクと同じタスクである TREC2003 Novelty Track 内の Relevant 文を見つけるタスクにおいて全体で 5 番目の成績を残した．

**SVDPACK http://www.netlib.org/svdpack/

*†TREC(Text REtrieval Conference)とは，テキスト検索方法論の大規模評価に必要なインフラストラクチャを提供し，情報検索コミュニティー内の研究の支援を目的とするワークショップ http://trec.nist.gov/

6.2 拡張精度比較

次に，各概念拡張が実際に，入力文をどのように拡張するかの一例を示す．

> Havelange also said that there would be no change to the schedule of the 1998 World Cup in France, which is slated for June 10 through to July 12.

図 4 拡張される文（入力文）

表 2-1 入力文の拡張結果

tf-idf		クラスタリング（概念数 800）		クラスタリング（概念数 1600）		主成分分析	
value	word	value	Word	value	Word	value	word
5. 32366	slate	0. 016764	Soccer	0. 159535	Staf	0. 789359	oil
4. 1987	cup	0. 012109	Match	0. 14934	teacher	0. 744623	advertis
3. 16667	schedul	0. 011589	Cup	0. 105831	tens	0. 616478	forecast
2. 85566	1998	0. 009665	Leagu	0. 060776	toppl	0. 567355	incom
2. 85516	franc	0. 009518	Game	0. 04558	wound	0. 556794	loss
2. 49302	world	0. 008799	Play	0. 032318	spanish	0. 534105	earn
2. 44673	chang	0. 008551	World	0. 0282	scan	0. 530325	ga
2. 39785	juli	0. 008391	Win	0. 026103	rodrigo	0. 524516	contract
2. 23457	june	0. 008239	Beat	0. 023215	weaker	0. 52435	export
0	-	0. 00819	Team	0. 01855	sector	0. 496788	Growth

表 2-2 入力文の拡張結果（WordNet）

上位概念へ拡張		上位->下位概念へ拡張		上位->下位->排他的拡張	
value	word	Value	Word	Value	Word
2	cash	2. 02564	Metal	2. 02564	Metal
2	metal	2	Cash	2	Cash
1	alteration change	1. 16667	June	1. 4	change
1	tournament	1. 16667	July	1. 33333	No
1	european_ nation	1. 16239	No	1. 16667	June
1	france	1. 04348	10 ten	1. 16667	July
1	indiana	1. 04348	12 dozen twelve	1. 14286	change
1	author writer	1. 02632	Beryllium	1. 07692	agenda schedule
1	effect issue outcome result upshot	1	Action	1. 04348	10 ten
1	change	1	Clothing	1. 04348	12 dozen twelve

図 4 の文は，1998 年のサッカーワールドカップ，フランス大会の話題を扱ったニュース記事の一文である．

表 2-1 で tf-idf の列は図 4 の文の tf-idf 値の高いものから 10 単語を示している．その他の列は tf-idf 値をクラスタリング、主成分分析の各手法で拡張した結果の活性値の高いものから 10 単語を示している．表 2-2 は，tf-idf 値を WordNet の手法で拡張した結果の活性値の高いものから 10 単語を示している．表 2-1 の各単語はステミングされているので単語として完全なものになってはいない．

表 1 において，クラスタリングの概念数 800 の方法が，一番 F-measure の結果がよい．表 2-1 をみると，クラスタリングの概念数 800 で拡張された単語も，franc，cup，1998 といった単語から soccer，match，cup などワールドカップに関連する言葉が上位に連想されている．同じクラスタリングによるアプローチではあるが，概念数が 1600 の場合は拡張結果が悪い．概念数がより増えたため 800 の時に扱っていた概念よりも細かい概念を扱っていると思われたが,staff，teacher などのように”soccer”とはかけ離れた単語が活性化されている結果となった．

主成分分析と WordNet による拡張は，表 1 において，拡張をしていない tf-idf と比較しても F-measure が落ちている．どちらも Precison では tf-idf と差はないが，Recall が落ち込んでいる．トピックに適合する文を，満遍なく見つけられなかったということになる．表 2-1 において拡張された活性値の高い上位 10 単語も，advertise，income，contract となっており，スポーツという分野ではなく，政治経済に偏った結果となった．すべてのトピックにおいてこのような結果ではないが，主成分分析での TREC のタスクの分類性能が tf-idf より劣るのは，この結果から自然に導かれる結果といえる．その原因として考えられるのは，主成分自体がうまく求められなかったことが考えられる．本研究では,単語の相関行列に SVDPACKC での計算の都合上，共起マトリックスをさらにスパースにしたものを用いた．本来 SVD は，共起マトリクスから相関係数を求め，そのマトリクスを相関行列として特異値を計算するのが一般的な手法だが，用いたマトリクスでは，きちんとした相関が得られなかったことが原因と考えられる．

表 2-2 から，WordNet では，France，World，cup といった単語から”soccer”の概念を得ることができなかった．文を上位概念へ拡張させたとき上位に活性化した単語，France，tounament が最もよく”soccer”の概念を表現しているといえる．この例の場合，下位概念への拡張，さらに排他的な制御を行うことによって，France，tounament の概念の活性値は低くなり，schedule，june といった“時間”の概念が高くなった．

図 4 の入力文は，”時間”の概念も含んでいるといえる文である．クラスタリング，主成分分析では，そのような概念を表現する単語が上位に来ることはなかった．クラスタリングや主成分分析はコーパス内に存在していない概念を表現することができないが，WordNet が人手で作られたシソーラスであり，幅広い語を取り扱っているという強みを示す結果となった．現状では，スポーツの概念の活性値が低くなってしまっているので，下位概念，排他的な拡張の計算法に工夫が必要であると考えられる．

6.3 多義語認識の検証

多義語の認識について,2 で挙げた Java を例にとって検証を行う．排他的活性を用いて多義語の認識を検証する．以下の図 5 の文を WordNet の各手法で拡張した結果,活性値が高いものから 10 単語を表 3 に示す．

図 5 の文は，java が”coffee”の意味を持つ状況の文である．

> Java coffee started from the Dutch, who planted the first Arabica trees in Java early in coffee's history.

図 5 java の例文

表 3 java の例文の拡張結果

上位概念へ拡張		上位->下位概念へ拡張		上位->下位->排他的拡張	
Value	Word	value	Word	value	word
3	coffee java	3	beverage drink	3.14286	coffee java
3	beverage drink	2	Island	3	beverage drink
2	Java	2	Metal	2.05263	tree
2	Indiana	1.67391	Tree	2.04348	in inch
2	Island	1.59184	in inch	2.03846	Indiana
2	Tree	1.47274	Indiana	2.01333	Java
2	in inch	1.4258	First	2	island
2	Metal	1.2867	History	2	first
1	History	1	Continuum	2	metal
1	Point	1	Past yesterday	2	History

図 5 の文を上位概念への拡張することで, java の"coffee"と"island"の概念が認識された. "coffee java"が java の"coffee"の概念で, "java"が, java の"island"の概念である. ここで, java の"programing language"としての概念は, WordNet で選んだ約 1 万語の概念に選ばれていないため拡張によって認識されることはない. 下位概念に拡張することによって文に含まれる時間や場所といった概念の活性度が上位となった. さらに排他的な制御を行うことによって, 上位に"coffee java", "beverage drink"といった概念が認識された. "島"の概念は活性値が低くなり,"飲料,コーヒー"としての java が拡張によって高い活性値を得られることになる. 概念を認識する上で排他的な制御が有効な働きをすることが確認できる.

7. まとめ

状況依存的概念拡張モデルの構築手法について述べ,状況に応じた動的な概念表現拡張モデルの実現に 3 つの手法を提案, 比較, 検証を行った.

本研究に用いたデータの次元が大きいため, 主成分を分析するためには, 相関マトリクスをスパースにしなければならなかった. このため, きちんとした主成分が得られていないと思われる. 主成分分析を用いた概念表現では, 高次元データから主成分をいかに抽出するかが課題として残っている. WordNet の利用では拡張の計算法に問題は残ってはいるものの, 排他的拡張によって, 状況依存的概念表現は達成できた. また, クラスタリングによる概念表現では, 有効な状況依存的概念表現が実現でき, TREC の Relevant 文を見つけるタスクにおいて 5 位の結果を示す事ができた.

概念拡張において, 排他的な制御が有効であることはすでに述べた. この制御をクラスタリング, 主成分分析の中心概念のモデルに適用することによって, 拡張精度の向上があると考えている. その検証, 比較が今後の課題である.

参考文献

[Richardson, 1995] R.Richardson, A.Smeaton, Using WordNet in a Knowledge-Based Approach to Information Retrieval, Working paper CA-0395, School of Computer Applications, Trinity College Dublin, 1995

[Voorhees, 1993] E.Voorhees, Using WordNet to Disambiguate Word Senses for Text Retrieval, in Proceedings 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, pp. 171-180 1993

[Fukumoto, 2002] 福本文代, 鈴木良弥, WordNet の同義語クラスとその上位関係を利用した文書の自動分類, 情報処理学会論文誌, Vol. 43, No. 6, 2002

[Powell, 1985] M.J.D.Powell, Radial basis functions for multivariable interpolation, a review, in IMA Conference on Algorithms for the Approximation of Functions and Data, pp.143-167, RMCS, Shrivenham, 1985

[Micchell, 1986] C.A.Micchell, Interpolation of scattered data: distance matrices and conditionally positive definite functions, Construct. Approx., Vol.2, pp.11-22, 1986

[Broomhead, 1988] D.S.Broomhead, D.Lowe, Multivariable functional interpolation and adaptive networks, Conmlex Systems, Vol.2, pp.321-355, 1988

[Moody, 1989] J. Moody, C. Darken, Fast-learning in networks of locally-tuned processing units, Nural Computation, Vol.1, pp.281-294, 1989

[MacQueen, 1967] J. MacQueen, Some methods for classification and analysis of multivariate observation, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol.1, pp.281-297, 1967

[Amari, 1994] 甘利俊一, 酒田英夫(編), 脳とニューラルネット, 朝倉書店 , 1994

[Doya, 2002] 銅谷賢治, 伊藤浩之, 藤井宏, 塚田稔(編), 脳の情報表現-ニューロン・ネットワーク・数理モデル, 朝倉書店 2002

[Zechner, 1996] K. Zechner, Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), Copenhagen, Denmark, pp. 986-989, 1996

[Sakawa, 1997] 坂和和正,田中雅博, ニューロコンピューティング入門, 森北出版, 1997

[Tokunaga, 1999] 徳永健伸, 言語と計算 5 情報検索と言語処理, 東京大学出版会,1999

[Kita, 2002] 北研二, 津田和彦, 獅々堀正幹, 情報検索アルゴリズム,共立出版, 2002

[Shinnou, 2003] 新納浩幸, 佐々木稔, SVDPACKC とその語義判別問題への利用, 自然言語処理,Vol.10,No.2, pp.129-149, 2003