2008－NL－184 (4)
2008／3／27

VSMに基づくSVMと構文解析手法を用いた旅行案内システムの構築

劉 曄† 滕 智† 土屋 誠司†† 任 福継††

† 徳島大学大学院先端技術科学教育部

†† 徳島大学大学院ソシオテクノサイエンス研究部

770-8506 徳島県徳島市南常三島町2-1

E-mail:†{liuye,teng,tsuchiya,ren}@is.tokushima-u.ac.jp

あらまし Web上には多種多様な情報が莫大な量存在する現在，分からないことは何でもWebを利用して調べることができる．しかしその反面，情報過多な状況下で，本当に欲しい情報を見つけにくくなっていることもまた現実である．特に，検索したい情報に対する知識が乏しいとき（例えば，観光に出かける場合など）にはその現象は顕著である．そこで本研究では，ユーザが入力したクエリーをその内容によって事前にSVMにより分類し，その結果を利用してVSMにより類似度計算を行う構成の観光案内システムを提案した．実験の結果，約55%の性能であることが分かったが，SVMによるクエリー分類を行わない手法に比べて約22%性能が向上することを確認した．これらのことから，本研究で提案した構成は有効であるといえる．

キーワード 案内システム, SVM, VSM, 類似度計算

A Study of the Guidebook System with Application in SVM and Syntactic Analysis based on VSM

Ye LIU†, Zhi TENG†, Seiji TSUCHIYA†† and Fuji REN††

† Graguate School of Advanced Technology and Science, The University of Tokushima

†† Institute of Technology and Science, The University of Tokushima

2-1 Minamijosanjima, Tokushima 770-8506

E-mail:†{liuye,teng,tsuchiya,ren}@is.tokushima-u.ac.jp

Abstract Numerous various kinds of information on present Web, make the retrieval possible to everything exist. However, excessive information also make it difficult to retrieve the one you really want, particularly, when the knowledge is scarce (for instance, it goes sightseeing). Therefore, in our research, we propose a sightseeing guide system which classify the user's query with SVM and then calculate the similarity by VSM. In the experiment result, the accuracy was 55% and about 22% higher than the one without VSM classification. It can be said that the proposed system is effective.

Key words Guide system, SVM, VSM, similarity computing

1. はじめに

Web上には多種多様な情報が莫大な量存在する現在，分からないことは何でもWebを利用して調べることができる．しかしその反面，情報過多な状況下で，本当に欲しい情報を見つけにくくなっていることもまた現実である．特

に，検索したい情報に対する知識が乏しいとき（例えば，観光に出かける場合など）にはその現象は顕著である．

そこで我々は，コーパスに基づく旅行案内システムの構築を行っている[1]．このシステムでは，Web 上の観光情報を基に作成した観光案内のためのコーパスを用いている．コーパスには，各観光地の情報をひとまとめにして登録している．ユーザが入力したクエリーと観光案内コーパスとの類似度を算出することで検索を実現している．なお，類似度を計算する際には，構文解析を施し，必要な情報のみを抽出することで処理量を減らしている．しかしこのシステムでは，観光地の情報をひとまとめにしたため，ユーザが本当に望む情報ではないものも同時に出力してしまう問題があった．

そこで本研究では，観光案内コーパスの情報を細分化し，ユーザの望む情報を過不足なく提供できる観光案内システムの構築を目指す．また，ユーザが入力したクエリーと観光案内コーパスとの類似度計算を行う前に，クエリーを内容により分類することで，検索精度の向上を試みた．

2. システム構成

図１に本研究で構築した観光案内システムの構成を示す．ユーザが入力した質問（クエリー）を形態素解析する．この情報をユーザの質問のパターンを学習した SVM (Support Vector Machine)分類器にかけることで，入力クエリーを質問内容に応じて分類する．分類された質問内容の解が候補を Web から作成した観光案内コーパスから抽出し，VSM (Vector Space Model)により，入力クエリーと解候補のコーパスとの類似度を算出する．算出された類似度の最も高いコーパス文章を解とし，応答生成を行う．

3. コーパス

3.1 コーパスの構成

本研究では，観光情報として近畿地方の２府４県の観光地ならびに温泉地を対象とした．Web の情報[2][3]を参考に構築した観光案内コーパスの内訳を表１に示す．なお，観光案内コーパスはすべて手作業で作成している．

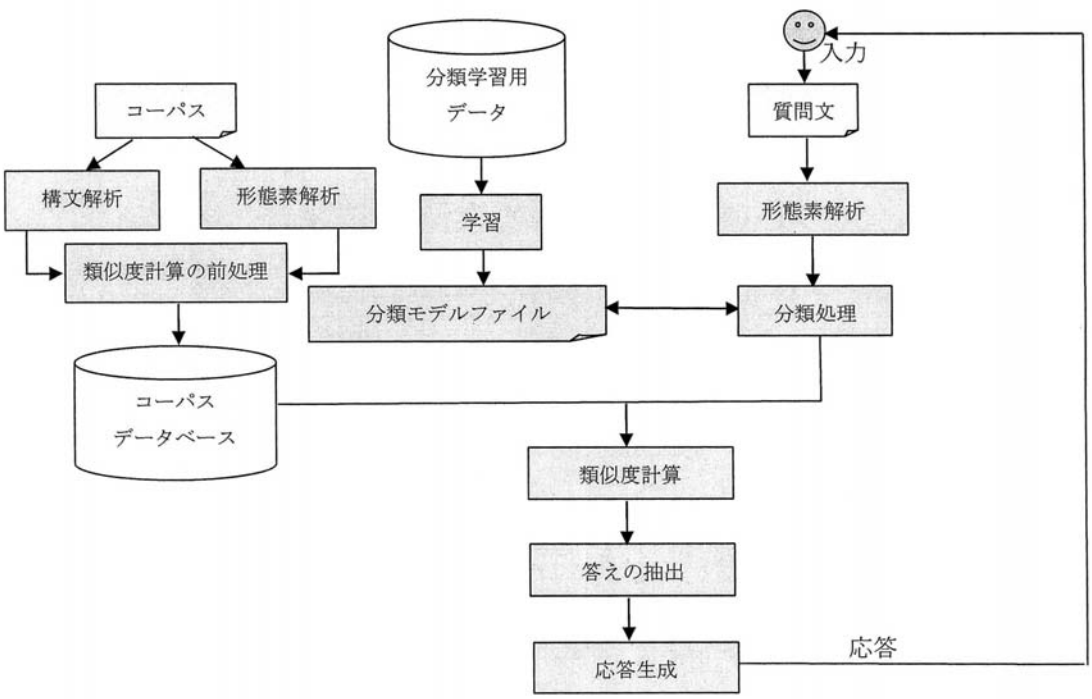

図１．システムの構成図

表 1.　コーパスの内訳

場所	観光地の数	温泉の数	合計
大阪	48	80	128
京都	45	6	51
兵庫	42	67	109
奈良	18	23	41
和歌山	30	34	64
滋賀	16	3	19
合計	199	213	412

3.2　格納方法

本研究では，入力クエリーとコーパスとの類似度計算にVSMを用いている．VSMでは，文章に含まれる単語をベクトルとし，ベクトル間の余弦により類似度を算出するため，文章に含まれる修飾語などの影響を受け大きく受ける．そのため，文章に含まれる単語のうち，その内容を表現するために重要となる単語のみを抽出し，ベクトル化することが検索精度向上に不可欠である．そこで，コーパスには，平文と共に構文解析を施した結果を格納している．

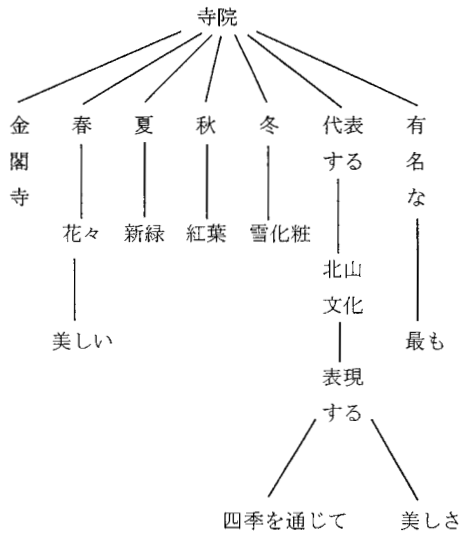

図 2.　構文解析結果の例

例えば，コーパスに格納する文章として，
「金閣寺は春は美しい花々，夏は新緑，秋は紅葉，冬は雪化粧と四季を通じてさまざまな美しさを表現する北山文化を代表する最も有名な寺院です．」
がある場合，この文章を構文解析する．構文解析した結果を図２に示す．ここで，
「金閣寺」，「春」，「夏」，「秋」，「冬」，「代表する」，「有名な」，「寺院」
という単語を抽出し，平文と共に登録する．このように処理することで，例えば入力クエリーとして
「春夏秋冬，最も有名で代表的なお寺はどんな寺院ですか？」
が与えられた場合，VSMを用いて精度良く適切な情報を検索することができる．

なお，構文解析には「Cabocha」[4]を利用した．

4.　クエリーの分類

4.1　分類処理

本研究では，入力クエリーの分類にSVMを使用した．SVMは数ある分類手法の中で最も分類性能の良い手法であると考えられている．SVMは，Vapnikらが1960年代に提案したOptimal Separating Hyperplane (OSH)が元になっている[5]．また，1990年代になってVapnik自身によって手法の拡張が行われ非線形識別が可能になっている．図３にSVMの基本的な分類原理を示す．空間上でクラス１とクラス-1を分類する際には，wを最小化することでマージンdが大きくなり，分類を実現できる．

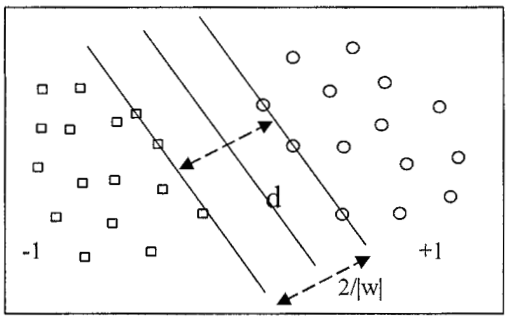

図３．SVMの概念図

SVMは基本的には２つのクラスを分類する手法であるため，複数のクラスを分類するためには他段階で分類器を用意し処理する必要がある．例えば，３つのクラスを分類する場合，始めに２つのクラスの分類を行った後，さらに一方のクラスを２つのクラスに分類することで３つのクラスの分類を実現する．

本研究では，SVMとして台湾国立大学のLinらによって作成されたライブラリであるLibSVM[6]を使用した．本研究で用いたパラメータの値を以下に示す．

```
-s svm_type:SVMタイプの指定(デフォルト0)
0--C-SVC
-t kernel_type:カーネル関数の指定
                              (デフォルト2)
0--線形(linear):u'*v
```

4.2 学習データ

SVMの学習に使用するデータには，20名の被験者からアンケートにより取得したデータを使用した．各被験者から５つずつ観光案内システムに入力するクエリーを取得した．取得したデータを人手により解析し，「場所」，「料金」，「時間」，「休業日」，「駐車場」，「アクセス」，「問い合わせ番号」，「特徴」，「まとめ」の９種類の質問タイプに分類した．なお，「特徴」とは，観光地についての簡単な紹介であり，「まとめ」とは，地域における観光地の簡単な紹介である．

5. 類似度計算

VSMはSaltonらにより提案された手法[7][8]であり，情報検索分野では広く利用されている．VSMによる検索では，高次元のベクトル空間上に配置した検索対象のベクトル表現と検索語のベクトル表現との相関量をその余弦によって算出する．算出式は以下のとおりである．なお，Dは文書を表し，文書Dに含まれる単語の重要度である重みがWである．

$$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}}$$

本研究では，文書中に含まれる単語のうち，構文解析の結果，重要であると判断された単語のみをベクトルの要素として利用している．また，その重みには出現頻度を利用している．構文解析により重要単語を抽出することにより，検索速度の向上と共に，検索精度の向上が期待できる．

6. 旅行案内システムの性能評価

6.1 評価データ

本研究で構築した旅行案内システムの性能を評価するため，18名の学生を被験者とする実験を行った．被験者から，本研究で定義した観光案内の分類９種類にたいする質問文を各分類について30問収集した．なお，本実験の被験者は，4.2節で説明したSVMの学習データを収集した被験者とは別の人物であり，本システムのコーパスの内容は一切知らない人物である．

また，収集した270問の質問には，本研究で構築したコーパスでは回答できない質問が約10%含まれていた．

6.1 システム全体の性能評価

表２．SVM+VSMの正解率

分類	データ数	正解	正解率
場所	30	5	16.67%
料金	30	19	63.33%
時間	30	23	76.67%
休業日	30	20	66.67%
駐車場	30	14	46.67%
アクセス	30	17	56.67%
問い合わせ	30	23	76.67%
特徴	30	24	80.00%
まとめ	30	4	13.33%
合計	270	149	55.19%

システム全体の性能評価を表2に示す.

全体の性能としては，約55%の正解率であり，特に，分類「場所」と「まとめ」の結果が著しく悪いことが分かる.

6.2 クエリー分類の性能評価

SVMの性能評価を表3に示す.

全体としては，約83%の正解率であるが，分類「場所」と「まとめ」の結果が著しく悪いことが分かる．これは，6.1節のシステム全体の評価と同じ傾向であり，クエリーの分類が悪影響を及ぼしているといえる.

これは，分類「場所」と「アクセス」，分類「まとめ」と「特徴」の情報が類似していることが分類性能の低下を引き起こす原因であると考えられる．実際，分類「場所」に分類されるべき質問が，分類「アクセス」に17問分類されおり，分類「まとめ」に分類されるべき質問が，分類「特徴」に19問分類されていた.

解決策としては，質問とその回答の内容がはっきりと分類できるようにする必要があると考えられる.

表3. SVMの正解率

分類	データ数	正解	正解率
場所	30	8	26.67%
料金	30	29	96.67%
時間	30	29	96.67%
休業日	30	30	100.00%
駐車場	30	30	100.00%
アクセス	30	30	100.00%
問い合わせ	30	30	100.00%
特徴	30	29	96.67%
まとめ	30	8	26.67%
合計	270	223	82.59%

6.3 類似度計算の性能評価

VSMの性能評価を表4に示す．なお，評価に用いたデータは6.2節のクエリー分類実験において正しく分類が成功したクエリーのみを対象として行っている.

全体としては，約67%の正解率であり，各分類において著しい性能の違いは見られない.

このことから，6.1節のシステム全体の性能にはVSMによる類似度計算よりもSVMによる入力クエリーの分類性能が大きく寄与していることが分かる.

表4. VSMの正解率

分類	データ数	正解	正解率
場所	8	5	62.50%
料金	29	19	65.52%
時間	29	23	79.31%
休業日	30	20	66.67%
駐車場	30	14	46.67%
アクセス	30	17	56.67%
問い合わせ	30	23	76.67%
特徴	29	24	82.76%
まとめ	8	4	50.00%
合計	223	149	66.81%

6.4 クエリー分類を行わない手法との性能比較

表6. VSMのみの正解率

分類	データ数	正解	正解率
場所	30	8	26.67%
料金	30	13	43.33%
時間	30	11	36.67%
休業日	30	7	23.33%
駐車場	30	11	36.67%
アクセス	30	6	20.00%
問い合わせ	30	16	53.33%
特徴	30	11	36.67%
まとめ	30	5	16.67%
合計	270	88	32.59%

本研究で提案したクエリー分類の有効性を示すため，クエリー分類を行わない手法との性能比較を行った．結果を表6に示す.

なお，比較実験に用いたコーパス及び入力クエリーは，6.1節で使用したものと同一であり，類似度計算の手法も6.3節のVSMと同様で

ある．つまり，本研究で使用したSVMによるクエリー分類のみを行わない手法と比較している．

全体としては，約33%の正解率があり，各分類において著しい性能の違いは見られない．

この結果から，本研究で導入したSVMによるクエリーの分類手法の有効性が確認できたといえる．

8. まとめ

本研究では，ユーザが入力したクエリーをその内容によって事前にSVMにより分類し，その結果を利用してVSMにより類似度計算を行う構成の観光案内システムを提案した．実験の結果，約55%の性能であることが分かったが，SVMによるクエリー分類を行わない手法に比べて約22%性能が向上することを確認した．これらのことから，本研究で提案した構成は有効であるといえる．また，同時に，分類性能がシステム全体に大きな影響を与えることも確認した．

これらの結果を踏まえて，今後，性能の良いクエリーの自動分類手法について研究していく予定である．

謝　辞

本研究の一部は科学研究費基盤研究B(課題番号1930029)の助成を受けて行われた．

参考文献

[1] Ye Liu, Zhi Teng, Fuji Ren, Shingo Kuroiwa and Seiji Tsuchiya, “Similarity Computing Based on VSM and Syntactic Analysis”, 2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, china, Aug.30-Sep.1, 2007.

[2] http://itp.ne.jp/contents/kansai

[3] http://michikene.ld.infoseek.co.jp/onsen.htm

[4] http://chasen.org/~taku/software/cabocha/

[5] V.Vapnik and A. Chervonenkis: ”A note one class of perceptrons.”; Automation and Remote Control, 25, 1964.

[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : “a library for support vector machines,” 2001.

[7] Salton, G., & Buckley, C. 1988. *Term weighting approaches in automatic text retrieval.* Information Processing and Management, 24(5), 513–523.

[8] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.