Query-Focused Multidocument Summarization Based on Hybrid Relevance Analysis and Surface Feature Salience

JEN-YUAN YEH[1], HAO-REN KE[2], WEI-PANG YANG[3]
[1]Dept. of Computer Science
National Chiao Tung University
Hsinchu 30010, TAIWAN
[2]Institute of Information Management
National Chiao Tung University
Hsinchu 30010, TAIWAN
[3]Dept. of Information Management
National Dong Hwa University
Hualien 97401, TAIWAN

Abstract: - Query-focused multidocument summarization is to synthesize from a set of topic-related documents a brief, well-organized, fluent summary for the purpose of answering an information need that cannot be met by just stating a name, date, quantity, etc. In this paper, the task is essentially treated as a sentence retrieval task. We propose a hybrid relevance analysis to evaluate the relevance of a sentence to the query. This is achieved by combining similarities computed from the vector space model and latent semantic analysis. Surface features are also examined to discern the impact of low-level features for query-focused multidocument summarization. In addition, a modified Maximal Marginal Relevance is proposed to reduce redundancy by taking into account shallow feature salience. The experimental results show the proposed method obtained competitive results when evaluated with the DUC 2005 corpus.

Key-Words: - Query-focused summarization; Hybrid relevance analysis; Latent semantic analysis; Modified Maximal marginal relevance;

1 Introduction

Recently, automated text summarization has drawn tremendous interest from both the natural language processing and the information retrieval communities. DUC (Document Understanding Conference) [8] is one of the active forums for evaluations of summarization systems. Since 2001, DUC has held several large-scale experiments for various types of summarizations; for example, single-document summarization, multi-document summarization, query-focused summarization. Query-focused multidocument summarization was first formally proposed at DUC 2005. The task is, given a set of topic-related documents, a query topic consisting of several complex questions, and a user preference profile[1], to generate a brief, well-organized fluent summary for the purpose of answering an information need. However, distinct from question-answering, the output summary can not just state a name, date, quantity, etc., which makes the task more challenging.

In this paper, we consider query-focused summarization as a query-biased sentence retrieval task. That is, only relevant sentences are included in the summary. The proposed method measures the relevance of a sentence to the query using a hybrid relevance analysis which linearly combines relevance measures obtained from the vector space model and latent semantic analysis. The output summary is generated by including sentences with the highest scores that are evaluated in terms of sentence relevance and low-level feature significance. A modified redundancy reduction module based on Maximal Marginal Relevance (MMR) [4] is proposed for anti-redundancy.

In the following, Section 2 introduces previous related work. An overview of the proposed system is presented in Section 3 and in Section 4 the proposed method is provided in details. Preliminary results are given in Section 5. Finally, Section 6 concludes this work.

[1] The user profile with a value of “general” or “specific” specifies the granularity required for the output summary.

2 Related Work

Much previous research has regarded query-focused summarization as a sentence extraction task which identifies sentences that are relevant to the query topic. For example, Daumé and Marcu [6] employed a Bayesian language model to estimate sentence relevance for ranking sentences. They showed that the Bayesian model consistently works well even when there is significantly less information. Bosma [3] created a graph representation for a document based on the Rhetorical Structure Theory (RST) [13]. A graph search algorithm is exploited to identify relevant sentences. Hovy *et al.* [10] proposed a method based on the extraction of basic elements. A basic element (BE), defined as a head-modifier-relation representation, is viewed as a basic unit to determine the salience of a sentence. Li *et al.* [11] built a query-oriented multidocument summarization system under the framework of MEAD [14] by integrating entity-based, pattern-based, term-based and semantic-based features. Ye *et al.* [17] handled query-focused summarization by computing sentence semantic similarity via concept links. Concept links were shown to outperform word co-occurrence since it highlights words that are semantically related. A modified MMR for anti-redundancy was proposed by introducing semantic similarity into the original MMR [4]. Schilder *et al.* [16] investigated a tree matching algorithm to obtain tree similarity of dependency parse trees between a question and candidate answer sentences. Sentences with the highest similarities are extracted as the summary. D'Avanzo and Magnini [5] exploited keyphrase extraction to identify relevant terms and used machine learning to select significant keyphrases. Summaries are generated according to relevance and coverage of keyphrases of a certain topic. Blair-Goldensohn [2] adapted a system originally designed to answer definitional and biographical questions and enhanced it with sophisticated question parsing, topic term identification, and passage retrieval. In addition, they experimented with several schemes for including the content of nearby sentence to help determine sentence relevance.

3 The Proposed System

An overview of the proposed system is presented in Fig. 1. The system first evaluates the relevance of each sentence to the query as well as its sentence significance on the basis of surface features; it then applies sentence selection by including sentences with the topmost scores to generate a summary of approximate 250 words to reflect the information need defined in the query topic and the level of granularity specified in the user profile.

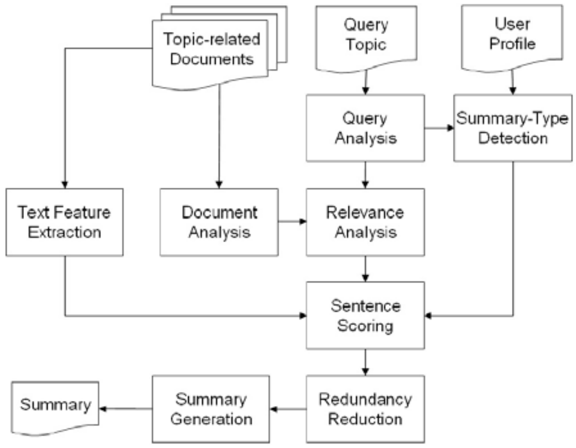

Fig. 1. System overview

There are eight modules in the proposed system:

1) Document Analysis preprocesses documents. It consists of sentence boundary detection, tokenization, Part-of-Speech (POS) tagging, stop-word removal, word stemming, and named entity extraction. After preprocessing, words tagged as NN (noun), VB (verb), JJ (adjective), or RB (adverb) are viewed as significant unigrams and are used to generate bigrams and trigrams. Bigrams and trigrams occurring in more than three sentences and all unigrams are kept to build a vector representation for each sentence using a term weighting scheme which was proposed by [1], as shown in Eq. (1). In this equation, N is the total number of sentences, sf_t is the number of sentence in which the word t appears, and $tf_{t,s}$ is the frequency of term t in a sentence s.

$$w(t,s) = \log(tf_{t,s}+1)\log(\frac{N+1}{0.5+sf_t}) \quad (1)$$

2) Text Feature Extraction (see Section 4.2) extracts surface features which are useful for query-focused summarization. These features are further exploited to measure the significance of a sentence in the sentence scoring module.

3) Query Analysis applies the same procedure of document analysis to the query and represents it as a weighted vector, except that the term weighting function uses Eq. (2).

$$w(t,q) = \log(tf_{t,q}+1) \quad (2)$$

where $tf_{t,q}$ the frequency of term t occurring in the query q.

4) Summary-Type Detection determines whether the desired summary is specific or general according to the user profile. If a specific summary is expected, the number of named entities in a sentence is combined into the sentence scoring function as an

additional feature.
5) Relevance Analysis (see Section 4.1) evaluates the relevance of a sentence to the query by measuring their similarity.
6) Sentence Scoring (see Section 4.3) combines sentence relevance and surface feature salience to estimate the significance of a sentence.
7) Redundancy Reduction (see Section 4.4) employs a modified MMR, which integrates feature salience into the original MMR [4], to reduce redundant information included in the summary.
8) Summary Generation (see Section 4.5) selects salient sentences on the basis of sentence score and re-orders sentences according to their original date of publication to form a summary.

4 Query-Focused Summarization

This section presents details for modules introduced in the previous section.

4.1 Relevance Analysis

For a sentence s, given a query q, the degree of the relevance of s to q is evaluated as their similarity, $sim(s, q)$. Three similarity metrics are introduced in this section.

4.1.1 Similarity Based on Vector Space Model

As shown in Eq. (3), $sim_1(s, q)$ is computed as the distance between the raw vector representations of s and q. This model has been proven successful in query-biased sentence retrieval [1] and is used in our work as a competitive baseline.

$$\begin{aligned} sim_1(s,q) &= \vec{s} \cdot \vec{q} \\ &= \sum_{t \in q} \log(tf_{t,q}+1)\log(tf_{t,s}+1)\log(\frac{N+1}{0.5+sf_t}) \end{aligned} \quad (3)$$

4.1.2 Similarity Based on Latent Semantic Analysis

Over the past few years, latent semantic analysis (LSA) [7] has been profitably employed in information retrieval. It has been shown the capability to derive inherent semantic structure from the corpus (i.e., the topic structure). We exploit LSA to measure the semantic relevance of a sentence to the query.

First, a *word-by-sentence* matrix A, presented in Eq. (4), is built. In the matrix, rows represent unique terms in the document collection and columns denote all sentences. An entry, $a_{i,j}$, is obtained by Eq. (1), which models the term weight of a term w_i in a sentence s_j. *Singular Value Decomposition* (SVD) is then performed on A. The SVD of A is defined as $A = UZV^T$ where U is an $m \times n$ matrix of left singular vectors, Z is an $n \times n$ matrix with a diagonal $(\sigma_1, \ldots, \sigma_n)^2$ and zeros elsewhere, and V is an $n \times n$ matrix of right singular vectors. Theoretically, Z could be interpreted as a *semantic space* (or the topic structure) derived from the corpus; U and V could be viewed as semantic representations of words and sentences in Z respectively.

$$A = \begin{array}{c|cccc} & s_1 & s_2 & \cdots & s_n \\ \hline w_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ w_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_m & a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{array} \quad (4)$$

Finally, *Dimension Reduction* is applied to Z by keeping only k ($k < r$) singular values to obtain an Z_k. By folding A into the reduced space Z_k using Eq. (5), a new matrix, $\widetilde{A}$, denoting the semantic representation of A in Z_k could be obtained.

$$\widetilde{A} = A^T U_k Z_k^{-1} \quad (5)$$

Similarly, for a query q=<$w_{q,1}$, ..., $w_{q,m}$>, it can be mapped into the same semantic space Z_k with Eq. (6).

$$\widetilde{q} = qU_k Z_k^{-1} \quad (6)$$

Thus, the semantic similarity between s and q can be measured by Eq. (7).

$$\begin{aligned} \widetilde{q} \cdot \widetilde{A}^T &= (qU_k Z_k^{-1})(A^T U_k Z_k^{-1})^T \\ &= qU_k (Z_k^{-1})^2 U_k^T A \\ &= < sim_2(s_1,q), \ldots, sim_2(s_n, q) > \end{aligned} \quad (7)$$

4.1.3 Hybrid Relevance Analysis

The hybrid relevance analysis is proposed to linearly combine $sim_1(s, q)$ and $sim_2(s, q)$ to obtain the effectiveness from both approaches. As a result, the hybrid similarity metric is defined as Eq. (8).

$$sim_3(s,q) = \alpha \cdot sim_1(s,q) + (1-\alpha) \cdot sim_2(s,q) \quad (8)$$

4.2 Surface Feature Extraction

In previous studies (e.g., [12], [18]), surface features, such as sentence position and tf-idf weighting, has been proven useful to identify significant sentences. In this work, we aim to understand whether these shallow features could still be evidences for sentence scoring in the query-focused summarization task. For a sentence s, we introduce the following surface features to obtain its feature scores:

f_1 – *position*: important sentences are usually

---

[2] If rank(A) = r, Z satisfies $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$.

located in the particular positions in a document. For example, in a news article, the first sentence always introduces the main topic. The position score is defined as Eq. (9) which was proposed by [9].

$$f_1(s) = 1 - \frac{NC(s)}{|D|} \tag{9}$$

where $|D|$ is the number of words in the document D that contains s and $NC(s)$ is the number of words appearing before s in D. Based on the scoring mechanism, the first sentence obtains the highest score and the last has the lowest score.

f_2 – *avg. tf-idf of significant words*: terms with higher term frequency and tf-idf values are more important. In order to obtain a more precise weight for s, we consider only significant words in s. The avg. tf-idf score is computed by Eq. (10) where $w(t, s)$ is the weight in Eq. (1).

$$f_2(s) = \underset{w \in s, w \in significant}{avg} w(t,s) \tag{10}$$

A significant word is defined as a keyword t which satisfies the criterion:

$$u + 0.5\sigma \le w(t,C) \tag{11}$$

where u is the mean, σ is the standard deviation of all $w(t, C)$, and $w(t, C)$ is the summation of all *tf-idf* values for t from all sentences in a document collection C.

f_3 – *similarity with title*: there is no doubts that the title always sums up main themes of a document. Therefore, the more similar the sentence is with the title, the more important it is. The similarity is obtained as Eq. (12) where s_{title} is the title sentence.

$$f_3(s) = sim(s, s_{title}) = \frac{\vec{s} \cdot \vec{s}_{title}}{|\vec{s}||\vec{s}_{title}|} \tag{12}$$

f_4 – *similarity with document centroid*: this measures the centrality of a sentence with the document. The centrality is considered as the similarity between s and the centroid of the document. More specifically, if a sentence contains more concepts identical to those of other sentences in the same document, it is more significant. This feature score is modeled as Eq. (13) where $D_{centroid}$ is the average of all sentence vectors in D.

$$f_4(s) = sim(s, D_{centroid}) = \frac{\vec{s} \cdot \vec{D}_{centroid}}{|\vec{s}||\vec{D}_{centroid}|} \tag{13}$$

f_5 – *similarity with topic centroid*: similar to f_4, this feature estimates the similarity of a sentence with the centroid of the topic cluster. The score is computed as Eq. (14) where $T_{cenroid}$ is the average of all sentence vectors in the document collection.

$$f_5(s) = sim(s, T_{centroid}) = \frac{\vec{s} \cdot \vec{T}_{centroid}}{|\vec{s}||\vec{T}_{centroid}|} \tag{14}$$

4.3 Sentence Scoring

The sentence score denotes the importance of a sentence and is used to determine salient sentences. We define the score of a sentence s in considerations of 1) its relevance to q, and 2) its salience of low-level features. Eq. (15) illustrates the score function.

$$score(s) = w_{sig} \cdot sig(s) + w_{sim} \cdot sim(s,q) \tag{15}$$

where $sig(s) = \sum_i w_{f_i} \cdot f_i(s)$, and $sim(s,q)$ could be any similarity metric proposed in Section 4.1. In $sig(s)$, $f_i(s)$ is one feature score in Section 4.2, and w_{f_i} is the weight for f_i used for linear combination.

Recall that in the system architecture, there is a module called summary-type detection. If the summary is determined to be specific, the number of named entities, $f_{NE}(s)$, is treated as another surface feature and is added to $sig(s)$. As a result, in this case, $sig(s) = \sum_i w_i \cdot f_i(s) + w_{NE} \cdot f_{NE}(s)$.

4.4 Redundancy Reduction

In [4], Maximal Marginal Relevance (MMR), presented in Eq.(16), was proposed for anti-redundancy in the multidocument summarization. The main idea is, when including a sentence s in the summary, it measures the MMR score of s to satisfy the following criteria: 1) the maximum relevance of s to q and 2) the minimum similarity to previously selected sentences which are already in the summary.

$$MMR = \arg\max_{s \in R-S} [\lambda sim(s,q) - (1-\lambda) \max_{s_i \in S} sim(s,s_i)] \tag{16}$$

where R is the ranked list of sentences, S is the set of selected sentences in the summary.

Since the original MMR only considers the similarity but no sentence representative power (e.g., surface feature scores), we propose a modified MMR which takes account shallow feature scores to address redundancy reduction. The modified MMR is shown in Eq. (17). In Eq. (17), δ and λ are weights to control the impact of $sig(s)$, SIM_1, and SIM_2, $sig(s)$, as mentioned in Section 4.3, is the score obtained from feature salience, SIM_1 denotes the similarity metric proposed in Section 4.1, and SIM_2 simply computes the cosine similarity.

$$MMR = \arg\max_{s \in R-S}[\delta \cdot sig(s) + \lambda \cdot SIM_1(s,q) - (1-\lambda) \cdot \max_{s_i \in S} SIM_2(s,s_i)] \quad (17)$$

In general, a sentence which has a high feature score and is highly relevant to the query but has a lower similarity to sentences in the summary will be ranked in the topmost position.

4.5 Summary Generation

After redundancy reduction, sentences are re-ranked according to new scores. The output summary is generated by selecting *k* sentences with the highest sentence scores. Finally, those selected sentences are presented in an order of the original date of their publication. Furthermore, if the length of the summary is greater than the required length (i.e., 250 words), the summary is truncated to satisfy the constraint. It is to be noted since we only focus on examining the effectiveness of hybrid relevance analysis and shallow feature salience, sentence ordering is not an issue discussed in this work.

5 Evaluation

In this section, we present our preliminary results.

5.1 The DUC 2005 Corpus

The DUC 2005 Corpus, consisting of 50 topics, was created by the NIST assessors. In each topic, there are approximate 25-50 documents. A user profile for each topic was also specified by the assessors to define the desired granularity of the summary. Then, other NIST assessors were each given a user profile, a DUC query topic, and a document cluster and were asked to create a summary that meets the needs expressed in the query and the user profile. These human-created summaries were used for evaluation as reference summaries.

5.2 Evaluation Metric

There were two evaluations conducted at DUC 2005. One automatically measured the consistence between the reference summaries and the machine-generated summaries using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [15]. The other was a manual evaluation to measure the linguistic well-formedness and the responsiveness of each submitted summary using a set of quality questions. Since we didn't participate in the DUC 2005, we only report here results evaluated by ROUGE.

ROUGE is an automatic evaluation tool for automated text summarization. It measures the number of overlapping units, such as *n-gram*, *word sequences*, and *word pairs* between the computer-generated summary and the ideal summaries created by human. There are several measurements: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. DUC 2005 ran ROUGE-1.5.5 and Jackknifing was implemented to compare human and system scores. Only the recalls of ROUGE-2 and ROUGE-SU4 were reported as the official ROUGE scores at DUC2005.

5.3 Results

In our experiments, we evaluated models with different settings (as listed in Table 1). Baselines 1-2 exploited the similarity metric proposed in [1] and the original MMR [4]; Model 6 is the system that integrates all proposed methods in this work. As for the parameters, they were set manually in these experiments: k=10 in Eq. (7); α=0.5 in Eq. (8); w_{sig}=0.5, w_{sim}=0.7, w_{f1}=0.8, w_{f2}=0.3, w_{f3}=0.3, w_{f4}=0.5, w_{f5}=1.0, w_{NE}=0.3 in Eq. (15); δ=0.5; λ=0.7 in Eq. (17).

Table 1: Settings of different evaluated models

Settings	Relevance	Features	MMR
Baseline 1	sim_1	N	None
Baseline 2	sim_1	N	Original
Model 1	sim_1	Y	None
Model 2	sim_1	Y	Modified
Model 3	sim_3	N	None
Model 4	sim_3	N	Original
Model 5	sim_3	Y	None
Model 6	sim_3	Y	Modified

The recall results are given in Table 2. In this table, the recall values of the best systems in DUC 2005 are also listed (see System 15 [17] and System 17 [11]).

First, Baseline 1 *vs.* Model 3 and Baseline 2 *vs.* Model 4 show the effects of different relevance analysis strategies no matter the modified MMR was applied or not. The result suggests that a hybrid relevance analysis which combines similarities computed from the vector space model and latent semantic analysis is a successful way to obtain a better sentence relevance to the query. Second, Baseline 1 *vs.* Model 1 and Model 3 *vs.* Model 5 imply that a scoring mechanism enhanced with surface features will improve the performance. Finally, Baseline 2 *vs.* Model 2 and Model 4 *vs.* Model 6 give an idea that the modified MMR which takes into account feature scores is a suitable for query-focused multidocument summarization.

To sum up, we got the best result of 0.075690 and 0.129950 for ROUGE-2 and ROUGE-SU4 respectively. The results are comparable to System 15 and System 17, which have the best results at DUC 2005.

Table 2: recalls of ROUGE-2 and ROUGE-SU4

Models	ROUGE-2	ROUGE-SU4
Baseline 1	0.064830	0.117550
Baseline 2	0.067690	0.120200
Model 1	0.069720	0.121350
Model 2	0.072280	0.124330
Model 3	0.068730	0.121750
Model 4	0.070110	0.124270
Model 5	0.073780	0.126870
Model 6	**0.075690**	**0.129950**
System 15	0.072510	0.131633
System 17	0.071741	0.129725

6 Conclusion

In this paper, we propose an approach to address query-focused multidocument summarization. The proposed method measures the relevance of a sentence to the query using a hybrid relevance analysis which linearly combines relevance measures from the vector space model and latent semantic analysis. Sentence salience is evaluated as well in terms of sentence relevance and low-level feature significances. Furthermore, a modified redundancy reduction module based on the original MMR is proposed for anti-redundancy. The experimental results show the proposed method obtained competitive results when evaluated with the DUC 2005 corpus. The contributions of this work are three-fold. First, a hybrid relevance analysis is proposed to estimate sentence relevance to the query. Second, shallow features are employed for scoring sentence importance and are shown to be useful. Finally, a modified MMR is proposed and examined to be a suitable component for query-focused summarization when surface feature salience is considered.

7 Acknowledgements

This work was supported by the National Science Council under grant number: 94-2213-E-009-013.

References:

[1] J. Allan, C. Wade, and A. Bolivar, Retrieval and Novelty Detection at the Sentence Level, *Proc. of SIGIR'03*, Toronto, Canada, 2003.

[2] S. Blair-Goldensohn, From Definitions to Complex Topics: Columbia University at DUC 2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[3] W. Bosma, Query-Based Summarization Using Rhetorical Structure Theory, *Proc. of CLIN 2003*, Belgium, 2003.

[4] J. Carbonell and J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, *Proc. of SIGIR'98*, Melbourne, Australia, 1998.

[5] E. D'Avanzo and B. Magnini, A Key-phrase-based Approach to Summarization: the LAKE System at DUC-2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[6] H. Daumé III and D. Marcu, Bayesian Query-Focused Summarization, *Proc. of COLING/ACL 2006*, Sydney, Australia, 2006.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, 1990, pp. 391-407.

[8] Document Understanding Conference. Available at http://duc.nist.gov/.

[9] T. Hirao, K. Takeuchi, H. Isozaki, Y. Sasaki and E. Maeda, NTT/NAIST's Text Summarization Systems for TSC-2, *Proc. of NTCIR 2003,* Japan, 2003.

[10] E. Hovy, C.-Y. Lin, and L. Zhou, A BE-based Multidocument Summarizer with Query Interpretation, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[11] W. Li, W. Li, B. Li, Q. Chen and M. Wu, The Hong Kong Polytechnic University at DUC2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[12] C. Y. Lin, Training a Selection Function for Extraction, *Proc. of CIKM'99*, Kansas, MO, 1999.

[13] W. C. Mann, and S. A. Thompson, Rhetorical Structure Theory: Toward a Function Theory of Text Organization, *Text*, Vol. 8, No. 3, 1988, pp. 243-281.

[14] MEAD, http://www.summarization.com/mead/.

[15] ROUGE, http://www.isi.edu/~cyl/ROUGE/.

[16] F. Schilder, A. McCulloh, B. T. McInnes, and A. Zhou, TLR at DUC: Tree Similarity, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[17] S. Ye, L. Qiu, T.-S. Chua and M.-Y. Kan, NUS at DUC 2005: Understanding Document via Concept Links, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[18] J. Y. Yeh, H. R. Ke, W. P. Yang, and I. H. Meng, Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis, *Information Processing & Management*, Vol. 41, No. 1, 2005, pp. 75-95.