AN EFFECTIVE QUESTION ANSWERING ENGINE BASED ON SEMANTIC THESAURUS[+]

I-Heng Meng[*], Wen-Tai Shieh and Wen-Chih Chen
Institute for Information Industry
Advanced e-Commerce Technology Laboratory, Taipei, Taiwan, R.O.C.
E-mail: < ihmeng, wentai, wjchen>@iii.org.tw

Abstract

Information retrieval has been proven effective for identifying specific passages within large volumes of data in response to a user query. However, users must click and browse numerous documents returned by keyword search to identify their desired word segments. The root problem is that keyword search is not an ideal method for users to express their real intentions for getting suitable documents. To overcome this problem, QA systems seek to process the question statement in natural language manner and find out the implicit intention of the user query. This study proposes a Chinese Question Answering system based on HowNet and Autotag to enable the system with semantic processing capability. The experiment collected 1000 Chinese News items from the ChinaTimes web site and presented an MRR (Mean Reciprocal Rank) value at 0.84.

Key Words: natural language processing, question and answering, concept similarity, passage retrieval

1. Introduction

Users must click and browse all documents returned by keyword search to identify their desired word segments. When numerous documents are returned, user time thus may be wasted as they must deal with many unsuitable documents. The root problem is that keyword search is not an ideal method for users to present their real intentions. To solve this problem, this work seeks to present queries in a natural language fashion, termed Question Answering or QA for abbreviation. It not only can fully describe the attributes of the user query, but also can express the real intentions of the user and help to retrieve answers more precisely.

A particular problem for Chinese question answering system is that groups of Chinese characters are not separated into meaningful words by blank spaces. This problem does not apply in English question answering systems and thus needs to be considered here. This study uses Autotag for separating Chinese characters into meaningful words. Autotag is a word segmentation system developed by Academia Sinica, Taipei. Meanwhile, HowNet (http://www.keenage.com) is considered the knowledge source for constructing our own thesaurus.

The main advantages of this work are described as follows. A robust natural language query parser is built for analyzing the query sentences to obtain the real intention of users. Because of the novel design, the system could comprehend the accurate meaning and get the keywords of a query. The system expands the query terms with their synonyms into a similar keywords set that is used as useful information for retrieving precise passages later. The second advantage is the provision of ability to extract Chinese semantic information from a large volume of documents and convert the extracted information into semi-structural XML files. The extracted semantic

+ A preliminary version of this paper appeared in The 2003 International Conference on Computational Science and Its Applications (ICCSA), May 2003, Montreal Canada, Lecture Notes in Computer Science, Vol.2667.
* Corresponding author

information related to concepts such as human, events, time, places and entities is used as the bases for passage retrieval. The third advantage is that a heuristic passage retrieval algorithm based on scoring method is established to get a higher precision rate. Especially, the significant breakthrough of this work is the ability to answer questions across different domains. Another difference between this paper and the preliminary version is the more concrete model about the calculation of concept similarity.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces the proposed question answering architecture. Section 4 then discusses the real implementation and experiment that yields a satisfactory rate of precision. Finally, Section 5 draws the conclusion and future research.

2. Related Work

Several well-known QA systems exist, such as LASSO [4], GURUQA [5], FAQ Finder [6] and AskJeeves [1]. LASSO interprets the questions entered by users with the NLP (Natural Language Processing) approach and groups them according to question type. After the intention of the question is extracted, LASSO identifies suitable answers from a large number of documents. The advantage of LASSO is that it extracts question types and intentions through an NLP approach to find more precise answers. This design thus differs from pattern matching with keywords. In GURUQA, a paragraph is the most fundamental processing unit. Every paragraph in a document is grouped by type and the paragraphs then are built into indexes. The same type of questions will match and find the answers through indexes. Finally, the results are listed from applying the algorithm for ranking the answers in terms of feature weights. Incidentally, GURUQA does not adopt the NLP approach. The FAQ Finder extracts information from questions such as the keynote of the question and keyword and uses this information to build indexes. The questions people asked are matched with these indexes to obtain possible answers according to the verbs and nouns in the question. Meanwhile, AskJeeves combines the keyword search with classification catalogues in different domains to find the required information using a concept-based approach. The above-mentioned systems generally extract types and focuses of questions that are used to match with the type of answer accordingly obtain more precise answers [4].

Most existing QA systems were designed for native English speakers and few Chinese QA systems exist currently. Some of the few existing work, like the ones developed by National Taiwan University (http://nlg3.csie.ntu.edu.tw) and Sinica Taipei (http://qa.iis.sinica.edu.tw), are primarily based on the corpus architecture. In addition to the above two sites in Taiwan, NKI (National Knowledge Infrastructure, http://www.nki.net.ch/User_handbook.htm) in China built a tremendous knowledge base that covers 16 specific domains, such as scientific, engineering, historical, medical, and so on, to provide the public with QA services. In this paper, we propose a quite different approach based on a frame model for processing Chinese natural language so as to provide semantic understanding capability, and this design has been proven effective and more accurate through an experimental evaluation.

3. Proposed Method

This study presents a more accurate result by means of recognizing concepts and relations within documents. Autotag is used to separate words and to recognize named entities in the question and Corpus. The similarity value of the relation between two concepts is calculated based on the definitions in the HowNet. The user question is parsed using Autotag first and then decomposed the words of the question to named entity and keywords. After keywords selection and query expansion, the passage retrieval module obtains suitable candidate passages segmented from these documents and ranks them by our scoring method. Figure 1 displays the system

architecture of our design.

The Question Parser module contains question named entity recognition and question type extraction. The processes of Document Parser are named entity recognition and decomposing corpus into sentences. Passage Retrieval module works based on the expanded query, the recognized named entity and ranking the candidates sentences by question type. The highest score of the candidate sentences will be chosen as the answer.

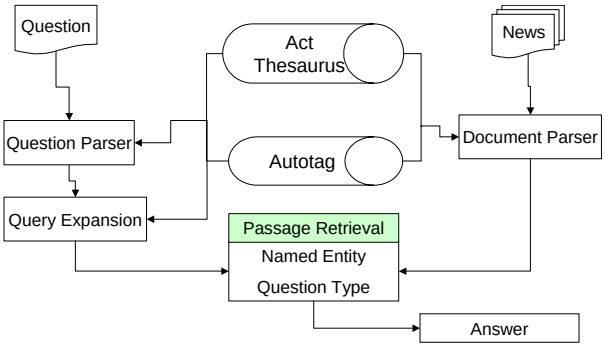

Fig. 1. System architecture of question answering engine.

3.1. ACT thesaurus

We construct ACT thesaurus based on the structure of HowNet to supply QA system with semantic knowledge. From the logical point of view, ACT thesaurus classifies Chinese vocabularies into three hierarchical levels, namely concept level, term level and sememe level. A concept is an entity that indicates a group of similar terms.

Definition 1: The *Sememe*(義元), denoted as *(S)*, is the finest granularity unit in ACT thesaurus.

Definition 2: The ACT thesaurus is composed of eleven sub-trees, namely Attribute Value Tree (T_{av}), Attribute Tree (T_a), Secondary Feature Tree (T_{sf}), Event Tree (T_{event}), Quantity Tree (T_q), Quantity Value Tree (T_{qv}), Entity Tree (T_{entity}), Syntax Tree (T_s), Converse Tree (T_c), Antonym Tree (T_{anto}), and EventRole Tree (T_{er}). The set of sub-trees, called *Knowmation Tree*, is denoted as $KT = \{T_{av}, T_a, T_{sf}, T_{event}, T_q, T_{qv}, T_{entity}, T_s, T_c, T_{anto}, T_{er}\}$.

Definition 3: The *Structural Term* in the ACT Thesaurus is defined as follows:

(a) A *Structural Term*, abbreviated as *ST*, can be denoted as (*T*).

(b) (*T*) = *ST* {*synset*} *DEF*, where *synset* indicates the collection of synonyms and *DEF* is the definition of *ST*.

(c) $DEF = \Lambda_{i=1\ to\ n}(S_i)$, where Λ indicates the 'concatenate' function. $(S_i) = EV_i|CV_i$, EV_i means the *English Value* of (S_i) and CV_i means the *Chinese Value* of (S_i). (e.g., affairs|事務)

(d) Each $(S_i) \in T_x$, where T_x is one of the sub-trees in *KT*.

Logically, a *Term* is mapped into a sememe set, called '*synset*'. *Synset*, including a number of items, is a collection of the synonymous sememes about the *Term*. After computing with the method introduced in later section, the item with a higher similarity value than the threshold is collected in *synset*. A structural term is notated with "*Term {synset} DEF*" and expressed in the form like "手機 {大哥大 手機} [用具 tool];<S>[交流 communicate]". It represents a term, called "手機 (cellular telephone)", has a *synset* with two synonyms, namely "大哥大(mobile phone)" and "手機(cellular telephone)". The *synset* is then followed by the definition of the term.

3.2. Concept Similarity

The similarity between two concepts (or called it *Terms*) is determined by using the definition of ACT Thesaurus. If two concepts are the same word or they have the same definitions in ACT Thesaurus, the similarity score will be set to 1. Otherwise, the similarity score is calculated based on the depth of the definition in the ACT Thesaurus structure tree. The following is the method used here for computing the similarity values between concepts.

(1) Those terms located on the same tree are

gathered into a set.

(2) Two terms with the same word definition or with the anti-meanings are chosen as the candidates for calculating their scores. For example, “貧富” vs “貧富” in Attribute tree, “窮” vs “富” in Attribute Value tree, and “良” vs “莠” in Secondary Feature tree.

Definition 4 (Level of a sememe): The *level L* of a sememe *(S)* in the sub-tree of ACT Thesaurus is denoted as *L(S)*.

Definition 5 (Common parents): The common parent *S* of sememe *Si* and *Sj* is denoted as $S^{(Si, Sj)}$.

Definition 6 (Score computing): Among the sememes belonged to the same sub-tree T_x, we choose two of them each time to calculate the score for the similarity computing purpose. The resulting *Score S’* of two Sememes is defined as $S'_{(Si, Sj)} = L_1 / (L_2+1)$, where S_i and S_j belong to the same type of sub-trees.

1. If $EV_i = EV_j$ and $CV_i = CV_j$, then L_1 and L_2 both represent the deepest level to which the sememes belonged to, and their values both are set to the maximum value of their levels, denoted as $max\{L(S_i), L(S_j)\}$.
2. if $EV_i \neq EV_j$ or $CV_i \neq CV_j$, let L_1 be $L(S^{(Si, Sj)})$ and L_2 is $max\{L(S_i), L(S_j)\}$. Furthermore, if S_i *and* S_j are recognized as antonym based on both T_c and T_{anto} in *KT*, then convert *S’* into a minus value.
3. $\forall$ *S’* $\in$ the same T_x, we get the most differentiated one, namely *S”*, to be the final score based on the following criteria.
   (1) If minus values exist then get the one with maximum absolute value in these minus values as *S”* regardless of the positive score values.
   (2) Otherwise, return the largest one as *S”*.

Definition 7 (Defined relational operations): Let $(T_i).Tree$ represent the set of trees that a term T_i contains.

- The Union operation is denoted as ($(T_i).Tree$ UNION $(T_j).Tree$).
- The Projection operation is defined by [$(T_i).Tree$].
- The Intersection operation is defined as ($(T_i).Tree$ INTERSECT $(T_j).Tree$)

Definition 8 (Concept similarity): We define the *Ratio* first. It is expressed with the following formula. The numerator indicates the number of trees with scores denoted as $\#T_{score}$. The denominator indicates the total number of trees.

$Ratio_{(Ti,\ Tj)}$ =COUNT([$(T_i).Tree$] INTERSECT [$(T_j).Tree$])/COUNT(($(T_i).Tree$ UNION $(T_j).Tree$).

Finally, the Similarity Value, abbreviated as *SV* and denoted as $Similarity_{(Ti,\ Tj)}$, of two terms is calculated with the following formula.

$Similarity_{(Ti, Tj)} = (\sum_{i=1\ to\ \#\ Tscore} S"_i)/\ \#T_{score} * Ratio + (Ratio*(1-Ratio))$.

3.3. Question Parser

The proposed design of Question Parser is to interpret the meaning and extract the intention from the asked question. We analyze the question sentence based on a semantic frame related to semantic concepts such as human, events, time, locations, and entities to extract valuable information. The format of the retrieved information, namely *Structured Question*, is defined in Definition 9.

Definition 9 (Structured question):

1. A *Structured Question*, abbreviated as *SQ*, is a structure resulted from parsing the question and it is denoted as *(Q)*.
2. *(Q)* = {*T, I, V,* S_1, S_2, O_1, O_2}, where *T* indicates the type of the question, *I* stands for the intention of the question. And *V,* S_1, S_2, O_1 and O_2 represent the verb(s), subject(s), subject modifier(s), object(s) and object modifier(s) parsed from the question, respectively.

The study investigates the usual patterns of Chinese question to classify them into three different categories, namely “特指問句 (Wh-Question)”, “選擇問句 (Disjunctive Question)”, and “是非問句(Yes/No Question)”, respectively. The question type *T* is assigned based on ACT thesaurus and the criteria mentioned in the following paragraphs. We build synonyms for these keywords appear in different types of question into ACT thesaurus in advance to find out the real intention of the question. For example, “何時(When)” and “幾時(When)” are synonyms and both indicate the inquiry for time.

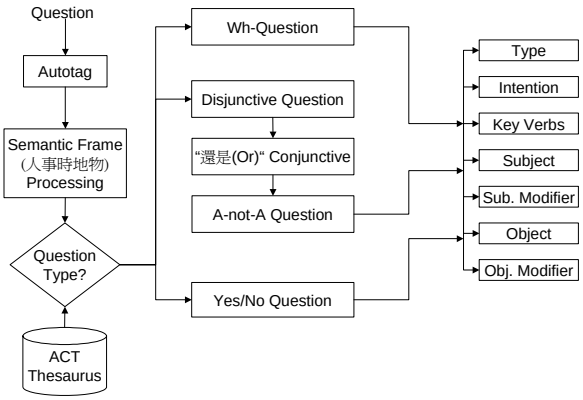

Fig. 2. Question parser process flow.

In Figure 2, the question sentence is parsed to obtain the part-of-speech information via Autotag. The nouns are then classified into either subjects or objects according to the question type and sequences of nouns. Based on the Chinese frame, the system judges which concept that the noun belongs to. The intention *I* of the question can be found from the question type combined with related nouns such as the subject and object.

3.4. Query Expansion

Our design expands the query by using the noun information in *(Q)*, such as S_1, S_2, O_1, and O_2, into a synonym set based on ACT thesaurus. The function described below gets similar concepts (terms) of a *Structural Term* named *(T)* from ACT thesaurus and Figure 3 presents the process of query expansion.

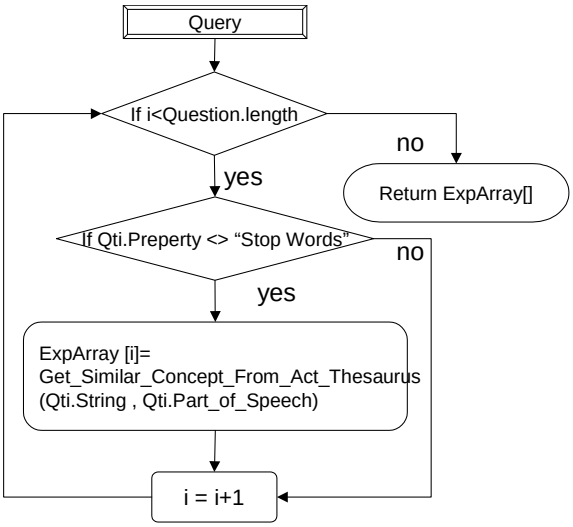

Fig. 3. The process flow of query expansion.

- Qt_i.String : *i-th* Token of the Query.
- Qt_i.Part_of_Speech: syntactical function of the *i-th* Token.
- Query = [Qt_1, Qt_2, Qt_3, …, Qt_m], where *m* denotes the *m-th* query terms.
- ExpArray[i] = [$Qt_{(i,1)} \mid SV_1$, $Qt_{(i,2)} \mid SV_2$,…, $Qt_{(i,n)} \mid SV_n$], where *(i, n)* means the *n-th* similar term of query tem i, and SV_n represents the similarity value of the *n-th* similar term as computing in Definition 8.

Get_Similar_Concept_From_Act_Thesaurus((T), (T).Part_of_Speech)

(1) For all synonyms in the *synset* of *(T)*, the similarity values with respect to *(T)* are calculated according to the equation described in Definition 8.

(2) The synonyms with values higher than the pre-setting threshold are chosen as the candidates for expanding the query. The returned synonyms are stored in an array fashion.

3.5. Named Entity Recognition

Autotag is applied to help us for segmenting a corpus into phrases and provides part-of-speech information. The named entity recognition model

will help us to decompose Human and Time concepts from the corpus and question. If a sentence in the corpus has a matching concept with the question, the sentence will get a higher score from the passage retrieval model.

人(Human) recognition. Chinese names comprise a one-character surname (or rarely, a two character surname) that is positioned before the given name, which also comprises one or two characters.

Theoretically, every Chinese person's name comprises a one or two character surname, followed by a one or two character given name. Thus the length of Chinese people's names ranges from 2 to 6 characters.

The following is the method used here for extracting Chinese names.

(1) Extracting the names of famous people using the "Nb" tag.

Many famous names are listed in the CKIP dictionary. The CKIP dictionary is issued by Academia Sinica, Taipei. If the proper noun starts with a surname listed in our surname table and the given name is more then one word, then it is considered a candidate for our lists. The precision of this step is over 95%.

(2) Extracting names using the surnames table and the scoring method.

The study designs a scoring method for dealing with non-striking names. Besides using surnames as clues for extracting names, titles and grammar rules are also used to help with name identification. An initial score value is set when a surname is identified in the corpus. The name is recognized if the score remains above the threshold value following the evaluation.

(3) Repair [2, 3]

Finally, a repair step is presented to increase the precision of the name extraction strategy. A person name may appear multiple times in a document. This study can repair candidates generated by step 3. The case listed below is used during repair.

Both C1C2C3 and C1C2 are included on the candidate list, and C1C2 is revised to C1C2C3.

事(Event) Relationship Extraction. The Event definition used in this study is the relationship among human, time, place/ organization, and entity. This definition is similar but less complicated than the Event/事件 class in HowNet. Consequently, the Event/事件 class and corpus are analyzed to obtain the three different type of relationships in our 事(Event) defined.

時(Time) Recognition. After decomposing the corpus, some clues for Chinese time extraction are identified.

(1) This study uses Autotag to help segment a corpus into phrases. Moreover Autotag is also introduced to provide part-of-speech information.

(2) From the analytical results, most Chinese times are consistent with the "Nd" tag.

(3) This study gathered eight Keywords that imply time， including "年(Year)", "月(Month)", "日(day)", "天(day)", "時(hour)", "點(clock)", "分(minute)" and "秒(second)".

(4) From the analytical results, Chinese times are consistent with seven kinds of tags. These tags are called legal tags ("Neu", "Nf", "Nes", "VCL", "Di", "FW", "D").

(5) The Chinese times can be classified into three types.

(a) absolute time: "民國九十一年五月二十八日(28 May, 2002)".

(b) relative time: "昨天(yesterday)", "當日(today)" etc. The relative time should be transformed into an absolute time. Two types of relative time exist, each with a different

translation rule.

(c) duration time: “九十一年四月至五月”. The duration time is concatenated two (時)time using the pattern with “(P)” tag.

3.6. Passage retrieval

The extracted information, namely *Structured Question*, acts as the input candidates for matching with the named entities extracted from a large volume of passages. A heuristic scoring method is applied to determine which passage with the highest score is most related to the question sentence. The similarity value *SV* between the query term and the candidate term is calculated in terms of the concept similarity algorithm mentioned in Definition 8. The following description is the scoring method used here for retrieving passages.

Scoring Method for Passage Retrieval

(1) A variable Qt_i represents the *i-th* query term of the query after query expansion. A variable Ct_j represents the *j-th* candidate term of the passage.

(2) The value of variable `Score` increases when the following conditions occur.

- If Qt_i and Ct_j have the same word definition, then `Score` advances 1.
- If Qt_i and Ct_j are synonyms then increase `Score` with their similarity value *SV*.
- If Qt_i belong to the concepts such as human, time, places and entities, then increase `Score` with its concept bonus. Each concept has its own bonus value setting.
- If Qt_i and Ct_j match successively, then add serial bonus to `Score`.

(3) Finally, if the candidate passage has the same intention with the question, then set the variable `Weight` to W, where W > 1, or assign 1 to `Weight`.

4. Implementation And Experiment

The real system named CNN (**C**hinese **N**ews a**N**swerer), as shown in Figure 4, is implemented for verifying our system architecture. The uppermost text box displayed in Figure 4 is designed for users to ask questions with respect to the fifteen different kinds of Chinese news as listed in the line below. The passage with the highest score is retrieved and displayed in the lower part of the screen. Then users could click the link named “全文(full text)” and show the full news content.

Fig. 4. The input question and the retrieved passage.

This experiment collects 1000 Chinese news items from 15 categories in the ChinaTimes web site (http://www.chinatimes.com). Ten pairs of questions and answers exist for each news item. The testing results are listed in Table 1.

Table. 1. The experiment results of MRR for each category.

Category	MRR
政治(Politics)	0.83
大陸(China)	0.86
股市(Stock)	0.84
社會(Society)	0.86
財經(Finances)	0.88
國際(International)	0.69
藝文(Art and Culture)	0.89
地方(Locality)	0.84
論壇(Forum)	0.86
影視(Entertainment)	0.81
運動(Sport)	0.85
開卷(Reading)	0.88
旅遊(Travel)	0.87
科技(Tech)	0.84
生活(Life)	0.85

From this result, we can observe that CNN is sufficient to cover different kinds of questions from users. And the precisions of these categories are averagely high except for the 'international' category. The reason comes from that there are too many foreign terms in the international news, we have not spent much of our efforts on this kind of contents for efficiency purpose yet. The excellent result indicates that the user is always satisfied with the first one answer returned by CNN. The QA engine proposed here has been included in the e-service system used by Advanced e-Commerce Technologies (ACT) laboratory in Institute for Information Industry (III).

5. Conclusions and Future Research

The QA system is a popular research issue in recent years. For the knowledge management level, how do we find the accurate answer from a large open domain corpus in a limited time is very useful for knowledge workers. In this paper, we propose a quite differentiated approach based on a frame model with processing Chinese natural language to provide semantic understanding capability and the design has been proven effective through an experimental evaluation.

Our research combines the technologies, including text mining, concept spaces, and construct a similarity concept table that includes synonymous and antonyms, together. These approaches help users to find exact answers efficiently, but there are some problems that should be improved:

(1) The precision of information extracted should be constantly promoted to obtain more accurate meaning of the question.

(2) The automatically ontology construction is still a research problem. A well-constructed ontology provides a rich semantic knowledge base for retrieving accurate answers.

(3) The proposed system does not deal with too much inference problems and the ability of inferences still stays in a research area. In many cases, there are different meanings and ambiguities when two words are combined and such situations appear frequently in the natural language area but the solutions for some difficult problems are still unclear.

Acknowledgement

This research was supported by the III Innovative and Prospective Technologies Project of Institute for Information Industry and sponsored by MOEA , Taiwan, R.O.C.

References

[1] AskJeeves , http://www.ask.com.

[2] H. H. Chen, Y. W. Ding and S. C. Tsai, "Proper Noun Extraction for Information Retrieval," *International Journal on Computer Processing of Oriental Languages*, Special Issue on Information Retrieval on Oriental Languages, 1998, pp. 75-85.

[3] H. H. Chen, S. J. Huang, Y. W. Ding and S. C. Tsai, "Proper Name Translation in Cross-Language Information Retrieval," *Proceeding of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, 1998, pp. 232-236.

[4] X. Li, and W. B. Croft, "Evaluating Question Answering Techniques in Chinese," presented as a poster at HLT 2001, in *Notebook Proceedings*, pp. 201-206.

[5] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju and V. Rus, "LASSO: A Tool for Surfing the Answer Net," *The Eighth Text REtrieval Conference (TREC 8)*, 1999, pp.175-184.

[6] J. Prager, E. Brown, A. Coden and D. Radev, "Question-answering by predictive annotation," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens Greece, July 2000, pp. 184-191.

[7] D. B. Robin, J. H. Kristian and A. K. Vladimir, "Question Answering from Frequently-Asked Question Files：Experiences with the FAQ Finder System," The University of Chicago Computer Science Department, *Technical Report TR-97-05*.