REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
10-01-2003

2. REPORT TYPE
Final technical report

3. DATES COVERED *(From – To)*
July 1, 2001- September 30, 2002

4. TITLE AND SUBTITLE

Automatic web searching and categorizing using query expansion and focusing

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-01-1-0917

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Sumali Conlon

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
School of Business Administration,
University of Mississippi, University, MS 38677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Navy Personnel Research Studies and Technology

10. SPONSORING/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING REPORT NUMBER

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We are in the process of build a prototype system that improves precision and recall rates for web search using query expansion and focusing techniques. We use linguistic analysis and co-occurrence information to analyze syntactic structures of the users' queries to improve search results.

One standard method of improving internet search is through query expansion. The major query expansion techniques add terms using (i) lexical semantic relations and (ii) relevance feed back. The lexical semantic relations in WordNet have been used widely as a main lexical resource for approach (i). Past research results indicate that using WordNet did not significantly improve information retrieval effectiveness. Our query expansion system also uses WordNet in a query expansion stage. However, instead of just adding all related terms from WordNet (synonyms, hypernyms, hyponyms, etc.) directly into user's queries, our system selects only useful additional terms. This selection process uses syntactic analysis combined with collocation and co-occurrence information from a large corpus collected from our domain of interest (IT).

15. SUBJECT TERMS

WWW, Information Retrieval, Query Expansion

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Sumali Conlon
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UU	5	**19b. TELEPHONE NUMBER** *(Include area code)* 662-915-5470

FINAL TECHNICAL REPORT

GRANT #: N00014-01-1-0917

PRINCIPAL INVESTIGATOR: Sumali Conlon (e-mail: sconlon@bus.olemiss.edu)

INSTITUTION: University of Mississippi

GRANT TITLE: Automatic web searching and categorizing using query expansion and focusing

AWARD PERIOD: July 1, 2001- September 30, 2002

OBJECTIVE: To build a prototype system that improves precision and recall rates for web search using query expansion and focusing techniques. We use linguistic analysis and co-occurrence information to analyze syntactic structures of the users' queries to improve search results.

APPROACH: One standard method of improving internet search is through query expansion. The major query expansion techniques add terms using (i) lexical semantic relations and (ii) relevance feed back. The lexical semantic relations in WordNet have been used widely as a main lexical resource for approach (i). Past research results indicate that using WordNet did not significantly improve information retrieval effectiveness. Our query expansion system also uses WordNet in a query expansion stage. However, instead of just adding all related terms from WordNet (synonyms, hypernyms, hyponyms, etc.) directly into user's queries, our system selects only useful additional terms. This selection process uses syntactic analysis combined with collocation and co-occurrence information from a large corpus collected from our domain of interest (information technology).

This work requires several steps, including:

1) Extracting noun and proper noun phrases from web documents
2) Collecting co-occurrence data from the web in the domain of interest.
3) Performing query expansion using information in the lexical database WordNet
4) Performing a focusing stage using co-occurrence information and syntactic analysis
5) Submiting the results to the web to retrieve additional web pages using the expanded phrases.

ACCOMPLISHMENTS (throughout award period):

We have completed many of the tasks listed above. However, the work is still ongoing since natural language processing requires an extremely sophisticated knowledge base and lexicon. The following describes each stage of the work which has been accomplished so far.

1) Extracting noun and proper noun phrases from web documents

This part benefits from the previous work supported by ONR. We selected proper noun phrases and other noun phrases semi-automatically from a KWIC (Key Words In Context) index file. The KWIC index file was created from data collected from the web in the information technology domain. This data set allows us to find many proper noun phrases. However, many proper nouns that are not found in the sources on the web were added by hand from other sources. Acronyms were also collected.

Noun phrases, proper noun phrases, and acronyms are important in internet search since most users' queries are short and are in the form of such expressions. If the queries are in the form of proper nouns or acronyms, the system identifies them from these extensive proper noun lexicons. They will not have to be expanded. The query "the office of naval research," for example, will not require the expansion stage. This query will be sent directly to the search engines (Google in our system). The returned URLs will be the URLs that the search engine provides.

2) Collecting co-occurrence data from the web in the domain of interest.

The KWIC program produces co-occurrence data that helps us in the query expansion and focusing stage. Here are some sample items in this file:

	W1	w2	w3	w4	w5	w6
1.	to	the	Cray-1	computer	which	was
2.	Apple	iMac	DV	computers	becomes	available
3.	ast-movingworld		of	computer	technology,	the
4.	immensely	powerful	IBM	computer	that	experts
5.	The	aster	PC	computer,	128+	MB
6.	answer	sessions	High-speed	computers	Quality	reference
7.	Cafe,	Polson,MT;	High-speed	computers,	fine	coffee,
8.	answer	sessions	High-speed	computers	Quality	reference
9.		for	high-speed	computers	and	electronics,
10.	techniques	for	high-speed	computers	have	developed
11.	satellites	and	high-speed	computers	and	}
12.	a	high-speed,	high-capacity	computer	containing	data
13.	Organizatin:Clientes		Fast	computer	systems	visit
14.	Possible	Tomorrows	High-speed	computers	have	made
15.	answer	sessions	High-speed	computers	Quality	reference
16.	of	a	powerful	computer	algorithm	written
17.	fast	and	powerful	computers	with	qually
18.	key	innovations	Powerful	computer	modeling	From
19.	turbidity	22 --	Powerful	computers	developed	by
20.	fast	and	powerful	computers	with	
21.	fast	and	powerful	computers	become,	they will

Table 1. Key Word In Context (KWIC) display for the sentences that contain the word "computer"

The actual data in the KWIC file contains 15 words per line. However, to make the output more readable in this report, we only show six words per line. Column 4 contains the word "computer" while the words around this column are words that appear before or after the word in question in the web documents we collected. Currently our KWIC file contains more than ten million records.

3) Performing standard query expansion using information in the lexical database WordNet

WordNet is a lexical database generated by a team of cognitive scientists at Princeton University. It is the most comprehensive lexical database available today and it has been used by most natural language processing researchers. In this research, we use entries in WordNet to perform query expansion.

Queries that are not proper nouns or acronyms may require query expansion. Users might submit queries that consist of a noun possibly modified by some adjectives ("big screen monitor," for example). Since there are many phrases that represent the same concepts, users' queries may not match the terms that are used by writers of web pages. The standard query expansion process is intended to fix this problem. The query "fast computer," for example, can be expanded based on the synonyms in WordNet, by finding the cross product of synonyms of each term in the query. In this example, "fast" has 74 synonyms:

accelerated	botonee	fast	hurried	locked	pernickety	scurrying	tinted
accelerating	botonnee	fastened	hurrying	meteoric	persnickety	secured	upright
alacritous	button-down	finical	immediate	meticulous	pinned	smooth	vertical
allegretto	choosey	finicky	immoral	moving	prestissimo	speeding	vivace
allegro	choosy	fixed	instant(a)	nice	presto	speedy	winged
andantino	constant	fleet	instantaneous	old-maidish	prissy	squeamish	
asleep(p)	dainty	fussy	invasive	old-womanish	prompt	stapled	
barred	double-quick	hastening	jet-propelled	overnice	quick	steady	
blistering	dyed	high-speed	knotted	particular	rapid	straightaway	
bolted	erect	hot	latched	pegged-down	red-hot	swift	

Table 2. Synonyms for "fast"

Similarly, the word "computer" has 9 synonyms:

Computer
computing machine
computing device
data processor
electronic computer
information processing system
calculator
reckoner
figurer
estimator

Table 2. Synonyms for "computer"

The query expansion stage will produce 74 x 10 = 740 phrases. Some examples are:

Accelerated Computer	Alacritous Computer	Allegretto Computer
Accelerated computing machine	Alacritous computing machine	Allegretto computing machine
Accelerated computing device	Alacritous computing device	Allegretto computing device
Accelerated data processor	Alacritous data processor	Allegretto data processor
Accelerated electronic computer	Alacritous electronic computer	Allegretto electronic computer
Accelerated information processing system	Alacritous information processing system	Allegretto information processing system
Accelerated calculator	Alacritous calculator	Allegretto calculator
Accelerated Reckoner	Alacritous reckoner	Allegretto reckoner
Accelerated figurer	Alacritous figurer	Allegretto figurer
Accelerated estimator	Alacritous estimator	Allegretto estimator

Table 3. Some phrases produced during the query expansion stage

The expanded queries may help improve the recall rate of a query since the additional terms might better match words in the web documents. However, most of the additional terms are not useful. As a result, if the system submits all of these phrases as additional queries, the precision rate will drop tremendously. In addition, search will become very slow, since each of the hundreds of queries will take several seconds to process. Thus, we must find ways to eliminate some useless phrases. The next step describe how we do this.

4) Improving on the standard method: focusing using co-occurrence information and syntactic analysis

The standard query expansion stage in step 3 helps us to find alternatives to the original query so that the search engine can find more pages that match the original query. As shown above, however, most of the expanded phrases obtained using the standard method are not useful.

Thus, in this stage we eliminate many of these useless phrases using a process we refer to as "focusing." This process narrows the set of expanded queries down to the most useful phrases, using co-occurrence data (see Table 1), together with syntactic analysis.

The main idea in his stage is that, if the user submits a query that is not a proper noun or an acronym, the system will try to find the phrases that represent the same concept as expressed by the user's query. It starts by expanding the original query by producing the cross product of the synonyms of each term in the query. The system then selects the useful phrases by learning from the previously collected web pages which combinations of synonyms make most sense. If the query is "fast computer," the system should produce additional queries like "high-speed computer," "high-speed parallel computer," "powerful computer," or "fast and powerful computer." However, phrases like "Accelerated Computer*," "Alacritous computing machine*," or "rapid growing computer company*."

To accomplish this the system performs two stages:

1) It uses the co-occurrence data to find the phrases that make most sense. In this example, the query "fast computer" may have "high-speed computer" as its synonym but not "accelerated computer." This is because the phrase "accelerated computer" never appears in the KWIC file, so the word "accelerated" must not make much sense in connection with "computer." In this step, we are able to eliminate many phrases produced by the previous stage.

2) The system performs syntactic analysis to find relevant phrases that are written differently from the expanded queries from the previous step (this work is currently ongoing). In addition to the phrase "high-speed" computer," for example, obtained from the original query "fast computer," there may be other phrases like "fast and powerful computer," or "high-speed parallel computer." To accomplish this stage, we use syntactic rules for noun phrases such as:

NP → N	computer
ART N	the computer
ADJ N	high-speed computer
ADJ ADJ N	high-speed parallel computer
ADJ ADJ ADJ N	fast parallel network computer
ART ADJ N	a high-speed computer
ART ADJ ADJ N	a fast parallel computer
N1 N2	network computer (N1 is a noun but serves as an adjective and N2 is a main noun)
N1 N2 N3	network IBM computer (N1, N2, and N3 are nouns but N1 and N2 serve as adjectives while N3 is a main noun)

This indicates that the synonyms of each word in the query can appear in many positions as long as they follow one of these rules. These rules also tell us that a phrase like "rapid growing computer company*" is about the "company" not the "computer" so the adjective "rapid" can not be used to modify "computer."

5) Submit the results to the web to retrieve additional web pages using the expanded phrases.

This stage uses the expanded phrases as search queries to send to the search engine. The retuned results are URLs that include results from the original query and the expanded queries.

CONCLUSIONS: Our query expansion techniques include two major stages: the standard expansion phase, and a new focusing phase, that selects among the expanded phrases to produce a subset of phrases that should make sense to ordinary language users.

SIGNIFICANCE: Though we have not yet been able to perform a systematic evaluation of our approach, our initial results show promise to improve precision and recall rates for internet search.

PATENT INFORMATION:

AWARD INFORMATION:

REFEREED PUBLICATIONS (for total award period):

1. Conlon, Sumali, John Conlon, and Tabitha James (*to appear*). "The Economics of Natural Language Interfaces: Natural Language Processing Technology as a Scarce Resource." *Decision Support Systems*, Elsevier Science Publishers, The Netherlands.

BOOK CHAPTERS, SUBMISSIONS, ABSTRACTS AND OTHER PUBLICATIONS (for total award period)

1. Conlon, Sumali J., Brad Babb, Barbara J. White (working paper), "Using Syntactic and Semantic Analysis to Perform Query Expansion for WWW Search." Target journal: ACM Transactions on Information Systems.
2. Conlon, Sumali, Jianfeng Wang, and Wendy Wang (2002). "Improving Internet Search using Query Expansion and Focusing." Proceedings of the Western Decision Sciences Institute Conference (DSI-2002), Las Vegas, NV, April.
3. Conlon, Sumali, Barbara J. White, and Wendy Wang (2002). "A Tool for Improving Knowledge Management." Presented at the Decision Sciences Institute Conference (DSI-2002), San Diego, CA, November (abstract).
4. Conlon, Sumali, Barbara J. White, and Wendy Wang (2002). "Finding Information on the Web using Linguistic Analysis." Presented at the Decision Sciences Institute Conference (DSI-2002), San Diego, CA, November (abstract).
5. Conlon, Sumali, Wendy Wang, Jie Zhang, and Yuehua She (2002). "Improving Internet Search using Query Expansion and Focusing." Presented at the Fifth International Military Applications Symposium Conference, June (abstract).
6. Wang, Wendy and Sumali Conlon (2002). "Automatic Extraction of Research Publications." Proceedings of the Western Decision Sciences Institute Conference (DSI-2002), Las Vegas, NV, April.
7. Wang, Wendy and Sumali Conlon (2002). "Reducing Information Overload using Automatic Text Summarization." Presented at the Decision Sciences Institute Conference, San Diego, CA, November (abstract).