Information Processing & Management Vol. 25, No. 3, pp. 265–273, 1989
Printed in Great Britain.

0306-4573/89 $3.00 + .00

AUTOMATIC THESAURUS CONSTRUCTION BY MACHINE LEARNING FROM RETRIEVAL SESSIONS*

U. GÜNTZER,† G. JÜTTNER,‡ G. SEEGMÜLLER,§ and F. SARRE†
†Department of Computer Science, Technical University of Munich, Postfach 202420, D-8000 München 2, FRG, ‡Institute for Application Oriented Knowledge Processing, Ulm, FRG, and §Gesellschaft für Mathematik und Datenverarbeitung, Bonn, FRG

(*Received* 27 *May* 1988; *accepted in final form* 15 *September* 1988)

Abstract—Users of information retrieval systems (IRS) know and use many relationships between concepts a long time before these find their way into textbooks, printed thesauri, or classification schemes. We present here an IRS component called TEGEN, which taps this expertise by automatically drawing conclusions from actual search behavior about possible thesaurus entries. This is done during an iterative knowledge acquisition process: only after explicit or implicit confirmation by other users of the IRS during the knowledge verification process, the results are incorporated into a thesaurus. TEGEN is written in PASCAL using a knowledge-based programming method. It uses the relational database system IMF2 and is implemented at the Technical University of Munich and at the Leibniz Computer Center of the Bavarian Academy of Sciences.

1. INTRODUCTION

Information retrieval systems (IRS) enable searches to be carried out on a collection of documents. During this process, the user receives information relevant to the questions that are of interest to him or her.

The information retrieval (IR) is more successful if various concept relationships can be incorporated into a search request by means of a thesaurus. Examples of such concept relationships are: synonyms, related terms, broader and narrower terms, and inflected forms.

Since the conventional manual methods of thesaurus generation are too costly and the automatic methods have not produced sufficiently good results to succeed in practice [1–3], we present in this article a thesaurus-generating system designated TEGEN, which provides a method by which a thesaurus can largely be learned by a computer without major effort or expenditure.

TEGEN is implemented on top of a multilingual IRS with automatic (and rather crude) indexing, which is used by professionals and graduate students. In such an environment a thesaurus is particularly helpful, namely:

1. The primitive kind of automatic indexing we use extracts all nonstop words from titles and abstracts of documents. This means that a paper on artificial intelligence might be indexed by the term "künstliche Intelligenz" if it happens to be written in German or by "AI" if only this abbreviation occurred in the title. Furthermore, a paper on trees might be indexed by "tree" or "trees" (or "arbres" etc.). Without the use of a thesaurus, recall will be rather low.
2. The thesaurus component is used interactively during query sessions. Thus, the user has immediate control when modifying the query with the help of the thesaurus.
3. Furthermore, in our environment—as in many others—gains in recall are much more at a premium than precision, because professionals are very good at screening out irrelevant material, but cannot afford to overlook relevant things; therefore, even if one would use the thesaurus only for *enlarging* queries (with a corresponding loss in precision) this would be beneficial.

*A shorter version of this article was presented at the conference "User-oriented Content-based Text and Image Handling" (RIAO88), MIT, Boston, MA, March 1988.

To summarize: If the indexing method is rather primitive, then a good thesaurus can make up for its deficiencies and achieves considerably better search results. But normally, if an IRS had adequate resources to create and maintain a good thesaurus, presumably it could also index more intelligently in the first place. The TEGEN system is intended to be a remedy to this dilemma. The intelligence of the user population exhibited during query sessions is tapped to construct a thesaurus reflecting *expertise*, *interests*, and even *jargon* of this population. Of course, it remains to be proven that a really good thesaurus can be generated this way.

2. THE BASIC PRINCIPLES OF THE LEARNING PROCESS

2.1 *Learning by analyzing*

The learning process used by the TEGEN system is called learning by analyzing. This is a refined variant of learning by observation in which, however, an additional feedback component is used to verify the learning results.

This learning process is based on the assumption that a large number of users of an IRS possess expert knowledge in the field to which their research relates. In carrying out their IR, they adopt certain procedures to find the desired documents contained in the collection of data.

The user searches are evaluated online to acquire information regarding significant concepts and their interrelationships simultaneously with the process of providing information concerning the relevant documents. Thus, the knowledge providing and the knowledge acquisition processes work cooperatively.

TEGEN observes the searches carried out by its users and exploits the users' expert knowledge. Probable semantic relationships among concepts are extracted online by means of acquisition rules from the syntax of a search request and from the responses of the user to certain reactions of the search system.

At the beginning of the learning process, the thesaurus is almost empty and can give only very limited support to an IRS; its quality, however, increases with the number of qualified users and instances of use.

2.2 *Feedback from the user*

User searches are evaluated online by TEGEN since feedback from the user is necessary for two reasons:

1. ambiguity of conclusions
2. uncertainty of results.

If the analysis of user searches is ambiguous as regards the conclusions to be drawn, the concept relationship is assigned by the user with the aid of explicit feedback. This means direct feedback from the user. The system asks questions to clarify the situation and expects these questions to be answered unambiguously.

Analysis of the search behavior of a user can lead to the acquisition of uncertain or wrong thesaurus contents. For this reason, the intermediate results of learning may be accepted as final only when their validity has been established by a sufficiently large number of verification processes on the basis of implicit or explicit feedback.

Implicit feedback is feedback from the user without direct questions. Thesaurus contents that require clarification are integrated in the search dialogue of the user at suitable points and the user's reactions to them are used for verification purposes.

2.3 *Architecture of the learning system*

Concept relationships, which are recognized as a result of the analysis of a user search, are designated *intermediate results* (IRES) and are initially assigned an appropriate status. These relationships have to be verified by further users so that their validity can be finally

established. A learning intermediate result becomes a *final result* (FRES) when the verification process is complete.

Thus, TEGEN consists essentially of two parts: knowledge acquisition and knowledge verification. The knowledge acquisition process derives possible thesaurus contents as intermediate results from the user searches. The knowledge verification process checks these intermediate results and converts them into final results. The architecture of the learning IRS is shown in Fig. 1.

The management of the IRES and the recording of the FRES is carried out by the thesaurus. This contains the relations SYNONYMS, RELATED TERMS, HOMONYMS, BASIC and INFLECTED FORMS, BROADER TERMS and NARROWER TERMS, NEGATIVE LIST and POSITIVE LIST, and USEFUL PREFIXES.

A few remarks concerning these relations shall clarify their semantics. The relation SYNONYMS contains not only synonym pairs in the linguistic sense but also translations and abbreviations, e.g. the pairs (artificial intelligence, AI) and (artificial intelligence, künstliche Intelligenz) would qualify for inclusions into the relation SYNONYMS. For multilingual information systems, the semiautomatic construction of translation tables (for at least those terms that are in actual use) is a great advantage.

The relation POSITIVE LIST contains the controlled vocabulary, i.e. those terms that have been used several times as search terms within queries. At a later stage we intend to use this list for indexing. The unary relation NEGATIVE LIST contains stop words like "is", "of", "the", and "der".

The relation USEFUL PREFIXES needs a special comment. It is designed to help the user when using truncation, e.g. "Datenba." shall denote all terms starting with "Datenba . . " like "Datenbank", "Datenbanken", etc. If the user truncates too early, too many documents will be derived; if the user truncates too late, recall will be lost. To find an adequate truncation point, the user goes through a term character-by-character from the left to the right and sends these prefixes one after the other as queries to the system. As an example, we display the results of this method for the term "Datenbanksystem" (this is the German word for "database system").

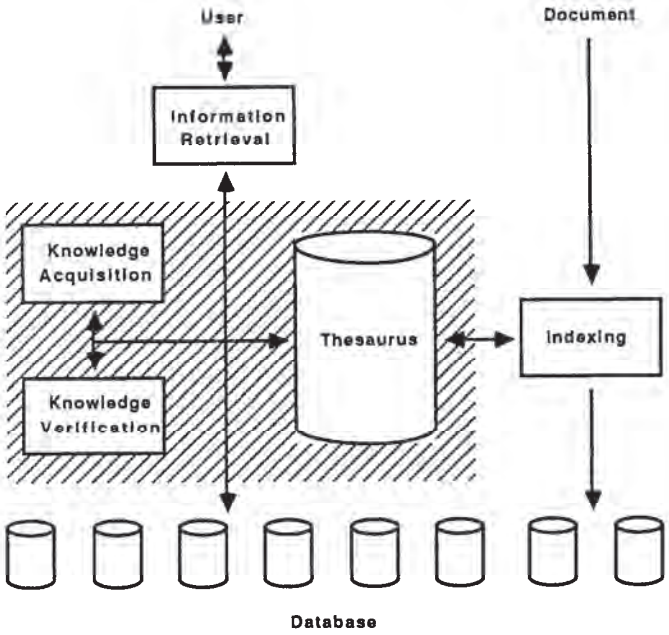

Fig. 1. Architecture of TEGEN.

Useful?	Prefixes	No. of documents retrieved
–	D.	52,578
–	Da.	17,050
no	Dat.	16,429
yes	**Date.**	2,182
no	Daten.	1,941
yes	**Datenb.**	523
no	Datenba.	469
no!	Datenban.	445
no	Datenbank.	445
yes	**Datenbanks.**	186
no	Datenbanksy.	162

As one can see, "Datenban." is not a useful prefix. If "Datenb." is not good enough (this depends on the rest of the query), then it will not be helpful to replace it by "Datenban." The user will go on and eventually use "Datenbanks." instead. Actually, the user will set the truncation point at those points, where a significant drop in the hit rate takes place or where the semantics changes, like between "Daten." and "Datenb."

3. KNOWLEDGE ACQUISITION

Acquisition of the intermediate results takes place by means of acquisition rules, which are represented in the form of production rules such as those frequently used for knowledge representation in expert systems [4–6].

The rules can be classified into three different types, which we present in the following sections. Examples of rules implemented in our system are given.

TEGEN currently contains 29 sets of semantic rules with a total of about 300 production rules that are implemented in PASCAL together with the rule interpreter. Experience gained in using the system will result in further acquisition rules.

3.1 *Indirect acquisition without feedback*

User searches allow an unambiguous conclusion to be drawn as regards a thesaurus content and this can accordingly be directly acquired without any query being directed at the user.

Insofar as rules of this class are used to investigate a similarity relationship, these cannot be specified more precisely as SYNONYMS, RELATED TERMS, or INFLECTED FORMS without feedback from the user. In this class of rules, therefore, similarity relationships are generally assigned to the relation RELATED TERMS and are specified more closely in a further stage of the process.

EXAMPLE.

Rule 14: Restriction by means of a new search request

Syntax:

if (a) two or more search terms x_i in a search request X are combined by OR
and
(b) x_i occur as key words
and
(c) the search request X is combined in a further search request Y by AND or AND NOT with further search terms y_j

then all x_i are accepted in pairs as learning IRES in the RELATED TERMS relation.

Semantics of Rule 14. If the user restricts the combination of search terms by the operators AND or AND NOT with further search terms in a further search request, the search terms of the first search request, which are used to form the union, in general

have a conceptual relationship. The fact that the OR combination receives further conceptual treatment by way of restriction is a strong indication that the terms x_i have not been combined accidentally.

Condition (b) of Rule 14 implies that the search terms are complete and sensible concepts even if truncated.

EXAMPLE.

X: (BAUM OR BAEUME OR TREE OR TREES)
Y: X AND (SUCHE OR SEARCH)
produces similarity relationships among:
BAUM, BAEUME, TREE, TREES.

3.2 *Indirect acquisition with feedback*

In this rule class, the conclusions regarding thesaurus contents are ambiguous; the situation is clarified by explicit feedback. Similarity relationships can be unambiguously assigned to the different thesaurus relations on the basis of queries addressed to the user.

EXAMPLE.

Rule 25: New attempt in the case of AND combinations
Syntax:

if (a) a search request X consists of n search terms x_i with $n \geq 2$
and
(b) the x_i are combined by AND
and
(c) the search request X is repeated in search request Y but one of the search terms involved, x_j, is replaced by the search term y
and
(d) the search request Y produces less results than X
and
(e) the user gets the results of Y printed
and
(f) . . . some further premise . . .

then (1) ask the user, whether:
(A) y is a narrower term of x_j
(B) y is a synonym or a translation of x_j
(C) x_j is an inflected form of y
(D) y is an inflected form of x_j
(E) an affinity relationship exists between x_j and y
(F) none of this applies
(2) in case (A)–(E), the pair (y,x_j) is accepted as IRES in the appropriate thesaurus relation.

Semantics of Rule 25. If, after an unsuccessful search attempt in which two or more search terms were combined by AND, a fresh attempt is made and one of the search terms x_j involved is replaced by a search term y, a conceptual relationship generally exists between x_j and y. If replacing x_j by y produces less results and the user is satisfied with the modification, because the results were printed, then it is rather likely that one has found in y a term narrower than x_j or related to x_j. This relationship is classified more precisely by queries addressed to the user.

The precondition $n \geq 2$ is necessary so that it can be seen that question Y is an emendation of question X.

EXAMPLE.

X: STORAGE METHOD AND INFORMATION RETRIEVAL
Y: INVERTED LIST AND INFORMATION RETRIEVAL
results, after query to user, in:
INVERTED LIST is a narrower term of STORAGE METHOD.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.