Timothy C. Craven
School of Library and Information Science,
The University of Western Ontario,
London, Ont., Canada

Customized Extracts Based on Boolean Queries and Sentence Dependency Structures

Craven, T.C.: **Customized extracts based on Boolean queries and sentence dependency structures.** Int. Classif. 16(1989)No.1, p.11-14, 27 refs.

A method is described for using Boolean queries in automatically deriving customized extracts from a text in which semantic dependencies between sentences have been coded. Each sentence in the structured text is treated as defining a separate extract. This extract consists of the sentence and all other sentences on which the sentence is directly or indirectly dependent for its meaning. Extracts from a text that satisfy a given Boolean query are merged to eliminate duplicate sentences.
A prototype implementation of the method has been developed within an experimental text structure management system (TEXNET). (Author)

1. Introduction

Different kinds of summaries of longer texts may be found useful by different kinds of users. Although production of different summaries of a single document to suit various user needs has traditionally been accomplished by hand (9), automated techniques have also been explored. In some approaches, the immediate source is the entire text of the original document (22) (23) (26). In other approaches, an intermediate document representation is created first, and it is from this intermediate representation that the various summaries displayed to users are in fact created (1) (10) (19) (21).

This paper assumes a fairly simple model to ensure comprehensibility of automatically customized summaries. Production of such a summary is viewed as a process of extraction of sentences from the source. Independent meaningfulness of the summary is ensured by making the extraction of a sentence conditional on the extraction of any other sentences required to provide the needed context. Retrieval of passages, in contrast to generation of new texts by automatic language processing, has been chosen because of the relative simplicity, economy, and efficiency of applying such techniques to a wide variety of databases, large and small. Sentences have initially been selected as the unit of retrieval in the belief that they represent a reasonable tradeoff between brevity and independence.

For the purpose of summary generation, the sentences of a text may be given an explicit dependency structure by human intervention (12) (19). Some work has also been carried out on automatic recognition of sentence dependency as well as of other semantic structures. Some approaches have involved text parsing, and knowledge representation in terms of frames of formal propositions (11) (13) (16) (27); other approaches work with more superficial analysis of the text (2) (3) (5) (15) (18) (20) (24) (26). Hybrid automatic/manual structure recognition is also possible (8). This paper simply assumes the structure as given, however derived in the first place.

Longer texts are likely to show fairly complex structures of sentence dependency. A sentence may be dependent for its meaning on an earlier sentence that does not necessarily immediately precede it. A sentence may in fact be dependent on two ore more earlier sentences, which are themselves not necessarily linked by any dependency (14). This has been shown to be true even of relatively short abstracts (17).

2. Boolean Qnery Extraction Model

The assumption that one or more sentences are to be extracted to provide context for other extracted sentences still leaves an unanswered question: how are the other sentences determined in the first place?

This paper assumes that the basis for the initial sentence extraction is a Boolean query, like those typically used in searching textual or bibliographic databases online.

Each sentence in a text may be seen as defining its own extract of the text. If sentence dependency is ignored, this extract consists quite simply of the sentence itself; for example,

The Baris subsequently were convicted and fined a total of $400 for violating the bylaw.

(Personal names in the illustrations have been disguised.)

Suppose, however, that sentence dependency is considered. The extract corresponding to a sentence may now be said to consist of that sentence plus all the other sentences on which it is directly or indirectly dependent for its meaning; for example,

The University Students' Council's plan to contest a controversial London housing bylaw won't be affected by a recent court decision upholding a similar regulation in Waterloo, says a USC official.

The *USC* has been monitoring *the Waterloo case,* where a bylaw restricts to five the number of unrelated persons in one house.

Ted and Barbara Bari, brother and sister, were charged after they rented rooms in a single house to eight students.

Waterloo launched an appeal after a Justice of the Peace initially ruled the bylaw discriminatory.

The Baris subsequently were convicted and fined a total of $400 for violating *the bylaw.*

Here, the middle three sentences are extracted because they clarify expressions in the original sentence: "the Baris", "subsequently", and "the bylaw". The original sentence is directly dependent on these sentences for its meaning. The first sentence is extracted because it in turn explains the meaning of the two expressions "the USC"

and "the Waterloo case". The original sentence is indirectly dependent on this sentence for its meaning. (At 5 sentences, the extract is still relatively small in comparison to the full document from which it is taken, which contains 24 sentences.)

The approach to extraction based on Boolean queries may be viewed as consisting of two stages: initially, the extract defined by each sentence is treated as though it were a separate retrievable document; then, if more than one extract is retrieved from the same document, these extracts are merged to eliminate duplicate sentences.

For example, take the following text, in which individual sentences have been labelled with capital letters:

A. Mark Sapher, Associate Director of Development at Western since March 1, 1986, has been promoted to Director of Development.

B. The appointment effective November 17, 1986, was announced by Angus V DeGide, Associate Vice-President University Relations and Development.

C. The Development Office reports to DeGide.

D. The University is preparing to announce a major fund-raising initiative in 1987, and DeGide and Sapher are working on plans for the size and direction of the campaign.

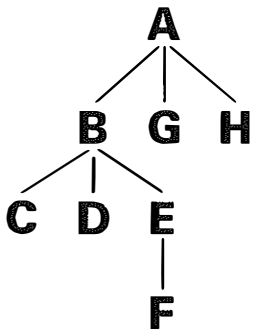
E. "As the University prepares to mount a major fund-raising campaign, we need to add key people to the Development Office", DeGide said.

F. "Mark Sapher has proven in the short time he has been at Western that he has a strong understanding of fund-raising in Canada and the Development Office at this University".

G. Western has begun to advertise for a new Associate Director of Development.:

H. Prior to this appointment at Western, Sapher had been Director of Advancement and Alumni Services at Lakehead University since 1982.

The sentence dependency structure can be represented briefly as:



A query "(mark sapher) and (lakehead university)" retrieves the extract corresponding to sentence H; that is it retrieves.

A. *Mark Sapher,* Associate Director of Development at Western since March 1, 1986, has been promoted to Director of Development.

H. Prior to this appointment at Western, *Sapher* had been Director of Advancement and Alumni Services at *Lakehead University* since 1982.

It should be noted that the personal name "Mark" does not appear in sentence H itself, but that is does appear in the extract as a whole, which therefore satisfies the query.

A query "degide and (fund raising)" retrieves three extracts: A + B + D, corresponding to sentence D; A + B + E, corresponding to sentence E; and A + B + E + F, corresponding to sentence F. When merged, these extracts form the single extract:

A. Mark Sapher, Associate Director of Development at Western since March 1, 1986, has been promoted to Director of Development.

B. The appointment effective November 17, 1986, was announced by Angus V *DeGide,* Associate Vice-President University Relations and Development.

D. The University is preparing to announce a major *fund-raising* initiative in 1987, and *DeGide* and Sapher are working on plans for the size and direction of the campaign.

E. "As the University prepares to mount a major *fund-raising* campaign, we need to add key people to the Development Office", *DeGide* said.

F. "Mark Sapher has proven in the short time he has been at Western that he has a strong understanding of fund-raising in Canada and the Development Office at this University".

The query "degide not said" retrieves

A. Mark Sapher, Associate Director of Development at Western since March 1, 1986, has been promoted to Director of Development.

B. The appointment effective November 17, 1986, was announced by Angus V *DeGide,* Associate Vice-President University Relations and Development.

C. The Development Office reports to *DeGide.*

D. The University is preparing to announce a major fund-raising initiative in 1987, and *DeGide* and Sapher are working on plans for the size and direction of the campaign.

3. Boolean Extraction from Less Structured Text

To the extracts just illustrated, we can contrast extracts based upon matching of query terms in the same paragraph. Note that every sentence in the sample text constitutes a separate paragraph except sentences B and C.

The query "(mark sapher) and (lakehead university)" yields an empty extract.

The query "degide and (fund raising)" produces

D. The University is preparing to announce a major *fund-raising* initiative in 1987, and *DeGide* and Sapher are working on plans for the size and direction of the campaign.

E. "As the University prepares to mount a major *fund-raising* campaign, we need to add key people to the Development Office", *DeGide* said.

Note the lack of explanation of who DeGide and Sapher are and what their exact relationship is.

The query "degide not said" generates

B. The appointment effective November 17, 1986, was announced by Angus V *DeGide,* Associate Vice-President University Relations and Development.

C. The Development Office reports to *DeGide*.

D. The University is preparing to announce a major fund-raising initiative in 1987, and *DeGide* and Sapher are working on plans for the size and direction of the campaign.

Observe how this does not directly identify what appointment is meant.

4. Prototype Implementation

As a prototype, the capability for such extraction based on Boolean queries has been implemented within TEXNET, an experimental text structure management package other features of which have been described elsewhere (4) (5) (6). TEXNET stores structured texts in a form developed earlier for linking anaphoric expressions to their antecedents (4). Editing software within TEXNET can be employed to display the structure and to make any modifications that may be required.

To generate a Boolean extract from the current structured text, the user keys control-B and then enters an appropriate Boolean query, which may include left or right truncated terms. The first up to 24 lines of the extract then appear on the screen. The user may scroll up and down through the entire extract at will by using the "PgUp" and "PgDn" keys. In addition the various other functions, the user may select to highlight the words in the extract display that match terms in the query; this is a feature commonly employed in retrieval systems to assist users in seeing how retrieval items respond to their queries.

Although results of the two-stage process for extraction outlined above are maintained by the prototype, the actual procedure followed is slightly different. Each sentence is examined for the occurrence of keywords from the query. In addition, it inherits keyword occurrences recorded for any sentence on which it is dependent. The result is an initial set of sentences, which is then augmented by the addition of any other sentences upon which these sentences are directly or indirectly dependent.

In a practical retrieval situation, Boolean-based extraction would most likely be a second stage following initial bibliographic retrieval. It would probably make use of the same query formulation. Thus, it would not necessarily require additional effort on the part of the searcher. One expection would be the use of the "not" operator, where it would be useful to be able to distinguish two cases: 1) not wanting to retrieve an extract where the document as a whole contains a given term; 2) not wanting the extract to include any sentences containing the term.

TEXNET includes software for automatically determining, in a rough way, the structure of a natural-language text. In terms of passage retrieval from the resulting structured text, this automatic analysis software is biased in favor of recall in preference to precision. In other words, its aim is more that sufficient sentences will be retrieved for comprehension than that all the sentences retrieved will be relevant to the reader's actual needs. This means that the typical structuring error is one of placing a sentence too far down the hierarchy.

Take, for example, the text of a typical obituary. When the automatic structural analysis option alone is used on this text, the Boolean query "victoria and hospital" yields the extract:

A. Dr Douglas Sanderson, Professor Emeritus of Surgery at Western, died October 8 in University *Hospital* at the age of 78.

B. Sanderson, who retired more than 10 years ago but continued to attend surgical rounds, was University *Hospital's* first chief of surgery.

C. Previously, he had been chief of surgery at *Victoria Hospital*.

D. Colleagues say that he made an indelible contribution to medicine at Western where he served as professor and head of the Department of Surgery for many years.

E. Many of the physicians now working at University and *Victoria* Hospitals were hired and trained by Sanderson.

Here, sentence D represents a failure of precision. If the same query is applied to a manually structured version of the same text, this sentence will not appear.

It may be worth noting that, in some cases, a document retrieved by a Boolean query could in fact yield an empty Boolean extract. For example, the structured document on Sapher's appointment above could be retrieved by the query "degide and (lakehead university)", but the extract based on the same query would be empty. For most purposes, such a document would best be treated as though it were not retrieved. The effect would be somewhat similar to that of proximity searching, though more sensitive to the actual semantic structure on the text.

Conversely, it should be noted that the extracts produced are not necessarily abstracts or even particularly short. If a large portion of the original document satisfies a given Boolean query, then a large portion is extracted.

Characteristic of certain kinds of stereotypic texts, including abstracts, is the use of cues or markers, such as "the hypothesis" or "is investigated", to indicate particular parts of the discourse (Liddy 1988; Radziewskaya, 1986). Such markers might be used in a Boolean extracting system to specify which aspects of a study are to be included in the extract. The texts used in the prototype implementation (articles from a university house organ) use a number of formulas, though not with perfect consistency. Authority for a new senior appointment is typically indicated by the formula "was approved by President . . . on behalf of the Board of Governors". Sentences indicating whom to contact for further information tend to contain: words like "contact"("-ing"), "information", "telephone" ("-ing"), "obtained" ("-ing"); the phrase "more information"; or the telephone exchange number "661".

5. Conclusion

This paper has outlined an approach using Boolean queries to derive customized extracts from structured texts. This approach has been implemented as a prototype within the TEXNET system. For practical use, however, it would most likely form part of a text retrieval system.

A number of other areas for possible further investiga-

tion remain. Textual units smaller than sentences might be experimented with. Somewhat deeper, or more formal, representations of the text might be used in generating the summaries. Users might be given options of specifying what types of contextual information they need. In addition to Boolean operators, automatic extraction might take account of other possible features of search specifications, such as proximity operators, role operators, and weighting. In addition to the present highlighting of search terms, other forms of enhancement might make clearer to users how the summaries respond to their queries.

Acknowledgement

Research reported in this paper was supported in part by individual operating grant A9228 of the Natural Sciences and Engineering Research Council of Canada.

Availability

A copy of the current version of the TEXNET software, which is written in Pascal for PC-DOS/MS-DOS, may be obtained by sending a 5 1/4" dual-density diskette to the author.

References

(1) Beardsley, B., and Cockburn, I.: ONTERIS printed index, vol. 1: abstracts of Ministry funded research and school board research in Metropolitan Toronto from 1959–1976. Toronto, Ont.: Ontario Ministry of Education, 1977.

(2) Berson, V.E., Blekhman, M.S., Zakharov, A.A., and Pevzner, P.P.: Computer implementation of a system analyzing supersentence connections. In: Automatic documentation and mathematical linguistics 18(1984)No.5, p.39-51.

(3) Blekhman, M.S.: The explicitness and implicitness of intersentence relations in scientific text. In: Automatic documentation and mathematical linguistics 18(1984)No.3, p.90-104.

(4) Craven, T.C.: A coding scheme as a basic for the production of customized abstracts. In: Journal of Information Science 13(1987)p.51-58.

(5) Craven, T.C.: Automatic recognition of sentence dependency structures. In: RIAO 88, p.911-924. Centre de Hautes Etudes Internationales d'Informatique Documentaire 1988.

(6) Craven, T.C.: Text network display editing with special reference to the production of customized abstracts. In: Canadian Journal of Information Science 13(1988)No.1/2, p.59-68.

(7) Craven, T.C.: Sentence dependency structures in abstracts. Accepted for publication in Library and Information Science Research.

(8) Craven, T.C.: An interactive aid for coding of sentence dependency structures. Submitted for the 1989 meeting of the Canadian Association for Information Science.

(9) Cremmins, E.T.: The art of abstracting. Philadelphia, Pa.: ISI Press; 1982.

(10) Derksen, D., Goodman, J.: ONTERIS: access to Ontario education information. In: American Society for Information Science Proceedings 24(1987)p.54-61.

(11) Fum, D., Guida, G., Tasso, C.: Forward and backward reasoning in automatic abstracting. In: COLING 82, p.83-88. Edited by J. Horecky. Amsterdam: North-Holland 1982.

(12) Girill, T.R.: Narration, hierarchy, and autonomy: the problem of online text structure. In: American Society for Information Science Proceedings 22(1985)p.354-357.

(13) Hobbs, J.R., Walker, D.E., Amsler, R.A.: Natural language access to structured text. In: COLING 82, p.127-132. Edited by J. Horecky. Amsterdam: North-Holland 1982.

(14) Hoey, M.: On the surface of discourse. London: Allen and Unwin 1983.

(15) Katzer, J., Bonzi, S., Liddy, E.: Impact of anaphoric resolution in information retrieval. Syracuse, N.Y.: Syracuse University School of Information Studies 1986.

(16) Kuhlen, R.: Some similarities and differences between intellectual and machine text understanding for the purpose of abstracting. In: Representation of knowledge as a basis of information processes: Proceedings of the 5th International Research Forum in Information Science. Amsterdam: Elsevier 1984. p.87-109.

(17) Liddy, E.D.: Structure of information in full-text abstracts. In: RIAO 88. Centre des Hautes Etudes Internationales d'Informatique Documentaire 1988. p.183-195.

(18) Liddy, E.D., Bonzi, S., Katzer, J., Oddy E.: A study of discourse anaphora in scientific abstracts. In: Journal of the American Society for Information Science 38(1987)No.4, p.255-261.

(19) Line, M.B.: Redesigning journal articles for on-line viewing. In: Hills, P.J. (Ed.): Trends in information transfer. Greenwood 1982. p.31-46.

(20) Mathis, B.A.: Improvement of automatic abstracts by the use of structural analysis. In: Journal of the American Society for Information Science 24(1973)No.2, p.101-109.

(21) Navy's ships will speak natural language. In: Bulletin of the American Society for Information Science 12(1986)No.4, p.12-13.

(22) O'Connor J.: Answer-passage retrieval by text searching. In: Journal of the American Society for Information Science 31(1980)p.227-238.

(23) Paice, C.D.: The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In: Oddy, R.N. et al (Eds.): Information retrieval research. London: Butterworths 1981. p.172-191.

(24) Polonskaya, O.R.: Logical-semantic connectors in the English language as formal indicators of text cohesion. In: Automatic documentation and mathematical linguistics 20(1986)No.3, p.91-94.

(25) Radziewskaya, T.V.: Texts of abstracts considered in a linguopragmatic aspect. In: Automatic documentation and mathematical linguistics 20(1986)No.4, p.55-63.

(26) Rush, J.E., Salvador, R., Zamora, A.: Automatic abstracting and indexing: II production of indicative abstracts by application of contextual inference and syntactic coherence criteria. In: Journal of the American Society for Information Science 22(1971)No.4, p.260-274.

(27) Taylor, S.L., Krulee, G.K.: Experiments with an automatic abstracting system. In: American Society for Information Science Proceedings 14(1977)p.83.

Prof. T.C. Craven, School of Library and Information Science, The University of Western Ontario, London, Ont. N6G 1H1, Canada