5. CHURCHMAN, C. W.: The nature of inquiry systems. New York, Wiley, 1969.

6. ROTHENBURG, D. H.: An efficiency model and a performance function for an information retrieval system = Information Storage and Retrieval, 5. köt. 3. sz. 1969. p. 109–122.

7. SCHUTZ, A.: Reflections on the problems of relevance. New Haven, Yale University Press, 1970.

8. SARACEVIC, T.: Relevance: A review of and a framework for thinking on the notion in information science = Journal of the American Society for Information Science, 26. köt. 6. sz. 1975. p. 321–343.

9. SARACEVIC, T.: id. mű.

10. CUADRA, C. A.–KATTER, R. V.: Opening the black box of „relevance" = Journal of Documentation, 23. köt. 4. sz. 1967. p. 291–303.

11. RATH, G. J.–RESNICK, A.–SAVAGE, T. R.: Comparison of four types of lexical indicators = American Documentation, 12. köt. 2. sz. 1961. p. 126–130.

12. O'CONNOR, J.: Relevance disagreements and unclear request forms = American Documentation, 18. köt. 3. sz. 1967. p. 165–177.

13. CUADRA, C. A.–KATTER, R. V.: id. mű.

14. REES, A. M.: Semantic factors, role indicators et alia: Eight years of information retrieval at Western Reserve University = Aslib Proceedings, 15. köt. 12. sz. 1963. p. 350–363.

15. CUADRA, C.: On the utility of the relevance concept. Santa Monica, CA. Systems Development Corporation, 1964.

16. REES, A. M.–SARACEVIC, T.: The measurability of relevance. Proceedings of the American Documentation Institute. 3. köt. Washington, D. C.. ADI, 1961. p. 254–334.

17. RESNICK, A.: Relative effectiveness of document titles and abstract for determining relevance of documents = Science, 134. köt. 3484. sz. 1961. p. 1004–1006.

18. FOSKETT, D. J.: A note on the concept of relevance = Information Storage and Retrieval, 8. köt. 2. sz. 1972. p. 77–78.

19. COOPER, W. S.: Utility-theoric versus relevance-theoric measures of effectiveness. Information Politics. Proceedings of the ASIS Annual Meeting. 13. köt. Washington, D. C. ASIS, 1976. p. 44.

20. SOERGEL, D.: Is user satisfaction a hobgoblin? = Journal of the American Society for Information Science, 27. köt. 4. sz. 1976. p. 256–259.

/REGAZZI, J. J.: Evaluating indexing systems: a review after Cranfield = The Indexer, 12. köt. 1. sz. 1980. p. 14–21./

(Novák István)

Automatikus eljárás tudományos és műszaki szakirodalmi dokumentumok szignifikáns szókapcsolatainak kiemelésére

1. Bevezetés

Az információkereső rendszerek egyik alapvető problémája az indexelés, a dokumentum eredeti információtartalmának igen tömör reprezentációja.

Ez a kísérleti rendszer a dokumentumok kivonatainak alapján igyekszik megoldani a feladatot. A kivonatok elemzése mellett szól, hogy reprezentálják a dokumentumok tartalmának elemeit, kiemelik a kutatási célokat, a módszereket, az eredményeket, a levonható következtetéseket stb., többnyire a dokumentumok szerzőitől származnak, a címbeli információval némely esetben kölcsönösen kiegészítik egymást, nem túl hosszúak, igen jellegzetes leíró stílusú mondatokat tartalmaznak és végül: *összefüggő szövegek, amelyek stiláris jellemzői kulcsot jelentenek a szavak és mondatok funkciója és ezáltal a kivonat tartalmának megértéséhez.*

Egy-egy szó, kifejezés vagy szókapcsolat szignifikáns voltának magából a szövegből kell meghatározhatónak lennie; ezért ez a megközelítés az eddigi módszereknél behatóbban vizsgálja *a kivonat szemantikai struktúráját.*

A szignifikáns szókapcsolatok kiemelésének módszere két előfeltevésen alapul, nevezetesen, hogy a fontos fogalmakat *nominalizált* (főneves alakra hozott) kifejezések jelölik, és hogy a szignifikáns szókapcsolatok és alkotóelemeik *különböző területeken más-más jelentésűek lehetnek.*

2. A főnévi szókapcsolatok kiemelése

A programrendszer öt modulból áll, ezek a következők.

2.1 Az input modul

Egy dokumentum input adatai *(1. ábra)* a következők:

egy azonosítószám és a cím,
a szerző(k),
a megjelenés helye és ideje – a folyóirat, a kötet, a szám és az év,
a kivonat szövege és
a kulcsszavak – ha vannak.

Az input modul a címet és a kivonat mondatait összekapcsolja és egyetlen karakterláncként továbbítja a következő modulhoz, a kulcsszavakat pedig úgy tárolja, hogy később majd összehasonlíthatók legyenek a rendszer által kiemelt szókapcsolatokkal. A többi adat változtatás nélkül, kártyakép formátumban egyenesen az output modulhoz kerül.

```
                    első oszlop
                        |
                        v
(1)  7323 SYSTEM ORGANIZATIONS FOR SPEECH UNDERSTANDING: IMPLICATIONS OF NETWORK AND
          MULTIPROCESSOR COMPUTER ARCHITECTURES FOR #AI
(2)  L. D. ERMAN, R. D. FENNELL, V. R. LESSER AND D. R. REDDY
(3)  * IJCAI•3•1973
(4)  (001) THIS PAPER CONSIDERS VARIOUS FACTORS AFFECTING SYSTEM ORGANIZATION FOR
           SPEECH UNDERSTANDING RESEARCH.
     (002) THE STRUCTURE OF THE HEARSAY SYSTEM BASED ON A SET OF COOPERATING,
           INDEPENDENT PROCESSES USING THE HYPOTHESIZE-AND-TEST PARADIGM IS
           PRESENTED.
     (003) DESIGN CONSIDERATIONS FOR THE EFFECTIVE USE OF MULTIPROCESSOR AND NETWORK
           ARCHITECTURES IN SPEECH UNDERSTANDING SYSTEMS ARE PRESENTED: CONTROL OF
           PROCESSES, INTERPROCESS COMMUNICATION AND DATA SHARING, RESOURCE
           ALLOCATION, AND DEBUGGING ARE DISCUSSED.
(5)  /*    SPEECH RECOGNITION, SPEECH UNDERSTANDING, SYSTEM ORGANIZATION, NETWORKS,
     /*    MULTIPROCESSORS, PARALLEL PROCESSING, REAL-TIME SYSTEMS, HARDWARE FOR AI,
     /*    SOFTWARE FOR AI.
```

1. ábra Egy dokumentum inputja

2.2 A szövegelemek kiválasztása

A címből és a kivonat modataiból kiválasztott, jelentős értelmű kifejezéseket *határolójelek választják külön.* A nem-alfanumerikus jelek és a szóközök mind határolójelnek számítanak, de az idézőjel, a kötőjel, a pont és a per-jel esetén a soron következő szimbólumot is vizsgálni kell: ha alfanumerikus karakter követi őket, nem kezelendők határolójelként, azaz a karakterlánc az illető ponton nem lesz elvágva.

A képleteket, képletszerű kifejezéseket és a tulajdonneveket ún. *diszkriminatív szimbólumok* beszúrásával ez a modul kizárja a további feldolgozásból. Az így előállított, kiemelt jelentéssel bíró szakkifejezéseket szignifikáns elemeknek tekinti a rendszer, és továbbítja a következő modulhoz.

2.3 A szakkifejezések ellenőrzése

Az ellenőrző modul a kiemelt szavakat sorra megkeresi a rendszer szótárában.

A szótárnak az a célja, hogy az adott szakterület fogalmainak ismerete és némi lexikális ismeret alapján lehetővé tegye a fontos szakkifejezések kiemelését. Nem lehet túl nagy (legfeljebb pl. 10 ezer szó), és biztosítani kell, hogy új szavakat is jelentős elemként lehessen kezelni. A szótár elemei a következő három kategória valamelyikébe tartoznak:

feltétel nélkül törlendő szavak, pl. stop-lista szavak, igék, határozószók stb.;

feltételesen törlendő szavak, pl. olyan melléknevek és főnevek, amelyek önállóan (izoláltan) nem szerepelhetnek szignifikáns elemként, ezek az ún. gyönge főnevek;

nem-törölhető, minden módosítástól is védett szavak.

A szótár jelenleg 2300 szót tartalmaz, kb. 49% az első kategóriába tartozik, és mindössze 1,2% a harmadikba. Minden szótári elem mellett jelezve van a kategóriája.

Az azonosított szavakat kategóriájuk szerint kezeli az ellenőrző modul. A védett szavakhoz nem nyúl, velük csak a következő modul fog tovább dolgozni. Ha egy elemet nem sikerült illeszteni egy szóhoz, a végződést kell ellenőrizni. Ha a végződés alapján sem sikerül az azonosítás, akkor az elem ún. *erős főnévnek minősül,* tehát olyannak, amely önmagában állva is szignifikáns elem lehet.

2.4 Szókapcsolatok generálása

Amikor ez a modul sorra veszi a beérkező szavakat, a fentiek értelmében már mindegyik be van sorolva négy kategóriá valamelyikébe; *törlendő szó – D, melléknévi jellegű szó – A, gyönge főnévi szó – W vagy erős főnévi szó – N.* A bejövő elemeket tehát sorra helyettesíteni lehet a fenti kategória-szimbólumok egyikével. Például:

```
"SEARCH STRATEGIES FOR THE TASK OF ORGANIC
   (N)      (W)      (D) (D)  (W)  (D)    (A)
CHEMICAL SYNTHESIS"
   (A)      (N)
```

A törlendő szavak, a melléknévi jellegű szavak és az izolált gyönge főnevek mintegy határoló jelekként funkcionálnak a szókapcsolatok kiemeléséhez. A fenti példajelsorozatból kiemelésre kerülő főnévi szókapcsolatok:

```
"SEARCH STRATEGY" és "ORGANIC CHEMICAL SYNTHESIS"
   (N W)                 (A A N)
```

2.5 Az output modul

A rendszernek kétféle outputja van. Az egyes dokumentumok önálló feldolgozásának eredményét a *2. ábra* illusztrálja. A dokumentumok egy-egy csoportjának feldolgozása után kerül sor az eredményül kapott kifejezések és szavak elemzéseire, erre példa a *3. ábra.*

```
(1) 7323   SYSTEM ORGANIZATIONS FOR SPEECH UNDERSTANDING: IMPLICATIONS OF NETWORK AND
           MULTIPROCESSOR COMPUTER ARCHITECTURES FOR @AI
              1 SYSTEM ORGANIZATION          2 SPEECH UNDERSTANDING
              3 NETWORK                      4 MULTIPROCESSOR COMPUTER ARCHITECTURE
              5 AI-@

          AUTHOR :  L. D. ERMAN, R. D. FENNELL, V. R. LESSER AND D. R. REDDY

(2)  001      1 SYSTEM ORGANIZATION          2 SPEECH UNDERSTANDING

     002      3 HEARSAY SYSTEM               4 HYPOTHESIZE-AND-TEST PARADIGM

            < WORD PROJECTION >
              1    1 : SYSTEM

     003      5 MULTIPROCESSOR               6 NETWORK ARCHITECTURE
              7 SPEECH UNDERSTANDING SYSTEM  8 DATA SHARING
              9 RESOURCE ALLOCATION         10 DEBUGGING

            < WORD PROJECTION >
              1    1 : SPEECH      1 : UNDERSTANDING      1 : SYSTEM
              2    1 : SYSTEM

(3)     KEY-PHRASE SET BY AUTHOR :
              1       SPEECH RECOGNITION
              2       SPEECH UNDERSTANDING
              3       SYSTEM ORGANIZATION
              4       NETWORKS
              5       MULTIPROCESSORS
              6       PARALLEL PROCESSING
              7       REAL-TIME SYSTEMS
              8       HARDWARE FOR AI
              9       SOFTWARE FOR AI

(4)     ABSTRACT PHRASE SET :
              1    1  DATA SHARING
              2    1  DEBUGGING
              3    1  HEARSAY SYSTEM
              4    1  HYPOTHESIZE-AND-TEST PARADIGM
              5    1  MULTIPROCESSOR
              6    1  NETWORK ARCHITECTURE
              7    1  RESOURCE ALLOCATION
              8    1  SPEECH UNDERSTANDING
              9    1  SPEECH UNDERSTANDING SYSTEM
             10    1  SYSTEM ORGANIZATION

(5)     PROJECTION WORD SET :
              1    1  SPEECH
              2    3  SYSTEM
              3    1  UNDERSTANDING
```

2. ábra Az 1. ábra adataihoz tartozó eredmények

```
KEY-PHRASES SET :  IJCAI-73 : BY PROJ. WORD - SEQ. :

    NO  PHRASE

     1  ABDUCTIVE REASONING PROCESS
     2  ABSTRACTION SPACE
     3  ABSTRACTION SPACE HIERARCHY
     4  ACOUSTIC ANALYSIS
     5  ACOUSTIC DATA
     6  ACTOR-@
     7  ADAPTIVE-CONTROL PROCEDURE
     8  ALGORITHM
     9  ANALYTICAL REASONING
    10  AND/OR GRAPH
    11  AND/OR TREE
    12  ARCHITECTURE
    13  ARITHMETIC
    14  ARTIFICIAL HAND
    15  ARTIFICIAL INTELLIGENCE
    16  ASSEMBLY MANIPULATION
    17  ASSERTIONAL INPUT SENTENCE
    18  AUTOMATIC PROGRAMMING SYSTEM
    19  AUTOMATIC PROTOCOL SYSTEM
    20  AUTOMATIC RECOGNITION
```

3. ábra Példa a 3. dokumentumcsoport kiemelt szókapcsolataira

3. A szignifikáns szókapcsolatok kiválasztása

3.1 A kivonat szövegének szerkezete – mondatközi kapcsolatok

A szövegbeli információ jelentős részének meghatározására – korábbi egyéb megközelítésekkel szemben – mi a szövegelemzés során a főnévi kifejezésekre koncentrálunk, ezeket tekintjük az információtartalom fő szemantikai hordozóinak.

Elemezni kell a mondatok közötti kapcsolatokat, minden egyes mondat szerepét a kivonat felépítésében és minden egyes főnévi szókapcsolat funkcióit a fentieken belül. Az elemzés *a főnévi szókapcsolatok közötti szemantikai összefüggések* által nyújtott információt, valamint a mondatokban található *szintaktikai és stilisztikai információkat használja fel.* A szemantikai és a szintaktikai feldolgozásnak szüksége van mind nyelvészeti ismeretekre, mind bizonyos pragmatikus tudásanyagra az adott szakterületen.

A mondatközi kapcsolatok nyomára vezető kulcsoknak több fajtája van, ilyenek pl. *a lexikai elemek különböző sémák szerinti párhuzamosságai, vagy ismétlődései, a kötőszavak, a mutatószavak.* Mi az ismétlődést alkalmazzuk. Elsősorban a mondatokban levő azonos elemek teljes ismétlődését vesszük tekintetbe, megengedve azonban néhány melléknév, de egy elem egy komponensének ismétlődését is, ha ez az ismétlődő komponens egy jelentős szótő. Ezeket a megismételt lexikai elemeket projektált (vetített) elemeknek nevezzük. Ha a mondatbeli főnévi szókapcsolatok egymásutánját tekintjük, ennek a felsorolásnak az első felében valahol van az „érvényes" ismétlődések – a projektált elemek – helye, tehát nem tekintjük a mondatkezdetet és a közvetlen környezetet. (A bonyolult vagy összetett mondatokat egyszerű mondatokká kellene és lehetne bontani, de a manuális beavatkozást igyekeztünk lehetőség szerint elkerülni, és csak eredeti mondatokat dolgoztunk fel.)

A szemantikai információ áramlásáról a következő feltevésünk van: az első kivételével a kivonat minden egyes mondata *elfogadja a „már adott" információt (témát) a megelőző mondatokból,* mégpedig a mondatbeli főnévi szókapcsolatok felsorolásának első felében. A főnévi szókapcsolatok felsorolásának *második felében füzi hozzá a mondat a korábbiakhoz az „új" információt.* Az utolsó kivételével a kivonat minden mondata átadja az információkat a következő mondatoknak.

3.2 A kísérletek eredményei

A fentiek értelmében a keresett szignifikáns szókapcsolatok kiemelhetők a főnévi szókapcsolatok halmazából. Azokat a szavakat, amelyek a fenti értelemben nem projektált elemek ugyan, de mind a címben, mind a kivonatban előfordulnak, járulékos projektált elemekként vesszük számba.

Öt dokumentumcsoport kísérleti feldolgozását végeztük el; ezek számítógéptudományi közlemények voltak, főként a mesterséges intelligencia kutatása köréből. Az összesen 346 dokumentumot tekintve a kivonatok átlagosan 5 mondatból álltak, a címek átlag 8, a kivonatok pedig átlag 107,8 szót tartalmaztak. A kivonatok szöveghossza alkalmas a gépi feldolgozásra. A dokumentumcsoportokra kapott eredményeket foglalja össze az *1. táblázat.*

1. táblázat

A szókapcsolatok/szavak átlagos száma

Elemek \ Dok. csoportok	1.	2.	3.	4.	5.	Összesen
1. Címek	1,9	2,0	2,0	2,1	1,9	2,0
4. Kivonatok	11,4	9,8	9,1	10,1	9,6	10,0
7. Szignifikáns szókapcsolatok	6,2	4,7	4,8	5,5	5,8	5,3
3. Szerzői kulcsszavak	6,2	5,9	6,5	4,0	5,4	6,0
8. Főnévi szókapcsolatok	7,4	5,1	4,7	5,7	5,9	5,5
5. Projektált szavak	3,3	2,3	2,2	2,7	2,8	2,6

A *1. táblázatban* szereplő sorszámok megfelelnek a 2. ábrában található sorszámoknak. A 2. ábra az 1. ábrán bemutatott input feldolgozásnak az eredményét közli. Jól összehasonlíthatók a szerző által megadott kulcsszavak (3), a címbeli főnévi szókapcsolatok (1) stb. A 3. ábra az egyik kísérleti dokumentumcsoport első húsz kiemelt szókapcsolatának az alfabetikus listája.

4. Következtetések és megjegyzések

A módszer számítógépes kísérletének eredményei alapján a következő megállapításokra jutunk.

(1) A szignifikáns szókapcsolatokat sikerült kiemelni a kivonatból. A cikkenként átlag kapott 5,3 szókapcsolat jól egyezik a szerzői kulcsszavak számával, de a két csoport között mégis adódtak eltérések, pl. az általánosság/specifikusság szempontjából.

(2) A kivonatból kiemelt főnévi szókapcsolatok száma átlagosan 10, mintegy 1,5-ször annyi, mint a szignifikáns szókapcsolatok, ill. a szerzői kulcsszavak száma.

(3) A címek önmagukban nem alkalmasak a dokumentumok szemantikai tartalmi elemzésére, figyelembevételük viszont indokolt, ezt bizonyítják a belőlük nyert járulékos szókapcsolatok. A címek átlagos szignifikáns szókapcsolatainak száma 2.

(4) A projektált elemek (átlagosan 2,6) valóban kulcsjellegűek, a szemantikai tartalom és a mondatok összekapcsolása tekintetében.

(5) Kvalitatív szempontból, a kivonatok szerzői stílussajátosságai folytán, az eredmények nem mindig egyértelműek. Egyre nő a kivonatok tartalmi és formai egységesítésének fontossága és szükségessége!

A fenti megközelítéssel olyan információábrázolásra törekednek, amely végső soron mondatfunkció – szignifikáns szókapcsolat-párok formájában valósulna meg. A kulcsszavazást az ilyen információábrázolás speciális esetének lehetne tekinteni, amikor csak a „téma" van megadva. Ami egy ilyen alapú információkereső rendszer felépítését illeti, további kutatásokat igényel (a) *a dokumentumok asszociatív tartalmi feltárása* – tehát nem egyszerűen a rokon tartalmú dokumentumokból való információfeltárás, hanem egyes dokumentumokban levő tartalmi asszociációt reprezentáló mondatok egymásnak megfeleltetésével, (b) *az automatikus tezauruszépítés* és (c) *az automatikus kivonatolás* kidolgozása.

/MAEDA, T. – MOMOUCHI, Y. – SAWAMURA, H.: An automatic method for extracting significant phrases in scientific or technical documents. = Information Processing and Management, 16. köt. 3. sz. 1980. p. 119–127./

(Szöllősy Éva)

ooo

Fogalmi kapcsolatok szemléltetése tezauruszokban

Az AIDOS–OS/ES programrendszer, valamint a megfelelő normatív dokumentumok és irányelvek alapján, továbbá a már meglévő tezauruszok alkalmazása során nyert tapasztalatok felhasználásával állították össze az NDK-ban *„Kisfeszültségű megszakítástechnika"* tezauruszát. A szerkesztési eljárásokat és módszereket részletesen leírták és szabályokban rögzítették. A tezaurusz szisztematikus részénél olyan megoldást alkalmaztak, amely lehetővé teszi valamennyi – az indexelés és visszakeresés szempontjából lényeges – fogalmi kapcsolat megkülönböztetését és áttekinthető ábrázolását.

1. Elméleti alapok

Elöljáróban a következő alapelveket szögezték le:

A deszkriptorok és nem-deszkriptorok – szavak és szócsoportok alakjában – *fogalmakat fejeznek ki.* Ugyanez vonatkozik a deszkriptorok és a több szóból álló kifejezések közötti kapcsolatokra is. (Ez utóbbiak nemcsak mondatok, hanem egyéb értelmes szerkezetek is lehetnek.)

A fogalmak (szemantikai megközelítésben) *szavakat és szócsoportokat jelentenek,* és (az ismeretelmélet tükrözési szabályainak megfelelően) az osztályok, az egyének gondolati visszatükröződései.

A fogalmi kapcsolatok így, egyrészt *a szavak és szócsoportok jelentései közötti kapcsolatok;* másrészt (ismeretelméleti szempontból) *az osztályok gondolati visszatükröződései közötti kapcsolatok.* A tezauruszban ez az egyes deszkriptorok közötti és az egyes nem-deszkriptorok közötti, valamint a deszkriptorok és nem-deszkriptorok közötti kapcsolatokban jut kifejezésre.

Az eddigi tezauruszoknál alkalmazott eljárásmódokkal ellentétben, az új tezaurusznál kizárólag az úgynevezett *egyértelmű ekvivalencia elve érvényesül.* Ennek előfeltétele azonban, hogy az AIDOS–OS/ES programrendszer inverz adatai, a deszkriptorok hivatkozási száma mellett, a nem-deszkriptorokat is tartalmazzák. A prekombinált, csoportosított megnevezések mellé pedig felvették a tezauruszba mindazokat a szerkezeteket is, amelyek – megadott kritériumok szerint – a deszkriptorok és nem-deszkriptorok kiválasztásához és meghatározásához szükségesek. Ezáltal, egyrészt a kifejezhetőség vált pontosabbá: a tezauruszban nem szereplő fogalmakat deszkriptor-kapcsolásokkal (posztkombinált megnevezésekkel) lehet kifejezni; más részről így biztosították, hogy *a tezaurusz a referátumok – későbbre tervezett – gépi indexelésénél is alkalmazható legyen.*

2. A fogalmi kapcsolatok szemléltetése és a gépi adatfeldolgozás

A tezaurusz szerkesztése közben arra törekedtek, hogy a fogalmi kapcsolatokat és, ha ezek anyagi jelenségeket tükröznek, akkor a fizikai jelenségek, a gyártmányok, a gyártási eljárások stb. közötti kapcsolatokat, a tezaurusz felépítése áttekinthetően szemléltesse.

A kapcsolatok ábrázolásának *gépi adatfeldolgozással történő megvalósítása esetén három lehetőség kínálkozik:*

utaló jelzetek használata;

a deszkriptoroknak és nem-deszkriptoroknak cikkekbe vagy szakcsoportokba való rendezése;

a deszkriptorok és nem-deszkriptorok – különböző részekre tagolt – hierarchikus beosztása és jobbra történő elcsúsztatása.

A betűrendes tezauruszok esetében kizárólag az elsőként említett megoldást alkalmazzák. Ebből következik, hogy a betűrendes tezauruszban nem valósítható meg optimális hatásfokkal valamennyi fogalmi kapcsolat átfogó kifejezése; különösen akkor nem, ha sok hierarchikus kapcsolatból tevődnek össze.

A cél olyan tezaurusz-forma létrehozása volt, amely lehetővé teszi *valamennyi leírandó fogalmi kapcsolat jól áttekinthető ábrázolását viszonylag kevés leírási elem*