

A PERMUTERM SUBJECT INDEX

Orosz Gábor

Országos Műszaki Könyvtár és Dokumentációs Központ

Az Institute for Scientific Information /Tudományos Tájékoztatási Intézet, Philadelphia, USA/ több évi kísérletezés után 1964-ben Science Citation Index elnevezéssel terjedelmes indexmű publikálását kezdte meg. Az indexmű többféle indexet ölel fel, amelyek egymáshoz kapcsolódnak és összefüggő indexrendszert képeznek. A Science Citation Index elnevezés nem szerencsés, mert a rendszer egyik részének jellegét tükrözi. Onnan származik, hogy a kiadvány eredetileg citation indexnek /irodalmi utalások indexének/ indult, s többi részei csak később alakultak ki.

A Science Citation Index az 1966-os évfolyamától kezdve a következő részekből tevődik össze:

Citation Index /irodalmi utalások indexe/: a dokumentumokban /folyóiratcikkekben, szabadalmi leírásokban/ szereplő irodalmi utalásoknak a szerzők neve szerint alfabetikusan rendezett és a forrásdokumentumokra utaló hivatkozásokat is feltüntető jegyzéke;

Source Index /források indexe/: a forrás-dokumentumok címadatait a szerzők neveinek betűrendjében tartalmazó bibliográfia;

Permuterm Subject Index /fogalompermutációs tárgyi index/: a forrás-dokumentumok permutált címszávaiból képzett betűrendes tárgyi index.

Az indexrendszer mindhárom része egy-egy esztendő kiválasztott folyóiratállományának cikkanyagáról, valamint kiválasztott szabadalmi leírásairól készül. Ennek folytán a forrás-dokumentumoknak ugyanaz a halmaza a három részlegben három különböző szempont szerint válik kereshetővé.

A Science Citation Index 1966-os évfolyama összesen 14 kötetet ölel fel. Ezekből a

Citation Index	5 kötet, az évfolyamban a Vol.1-5;
Source Index	3 " " " " " 6-8;
Permuterm Subject Index	6 " " " " " 9-14.

Az index-gyűjtemény kialakításának alapját az az elgondolás képezte, hogy az egész gyűjteményt, annak minden részét elektronikus adatfeldolgozó berendezéssel teljesen gépi uton, automatikusan állítsák elő abból az inputból, amit a segédszemélyzet az eredeti dokumentumok nyomán gépi adathordozó közegen /gépi lyukkártyán/ létesít. Az input a feldolgozott dokumentumokról kizárólag formális adatokat tartalmaz: a forrás-dokumentumok és az azokban irodalmi utalásként szereplő hivatkozott dokumentumok címadatait. Az indexek tehát indexelő szakemberek intellektuális közreműködése nélkül készülnek.

Az alábbiakban a Permuterm Subject Index első évfolyamával foglalkozunk, melyet a Science Citation Index 1966-os évfolyamának kiegészítéseként pótlólag adtak ki 1967 második felében.

A PERMUTERM SUBJECT INDEX RENDELTETÉSE

A Permuterm Subject Index a tárgyszavak szerinti keresés eszköze. A forrás-dokumentumok címeiben lévő információtartalmat tárja fel olyan formában, mely a cím bármelyik érdelemes szava nyomán végzendő keresésre biztosít lehetőséget.

Az index kettős feladat ellátására hivatott. Egyrészt arra, hogy ha valaki egy témához keres szakirodalmat, az a téma megnevezésében szereplő szavak bármelyikének nyomán eljuthasson mindazokhoz a dokumentumokhoz, amelyeknek címeiben a szavak előfordulnak. Másik rendeltetése, hogy abban az esetben is eszközként szolgáljon valamelyik dokumentum kikereséséhez, amikor az informálódó a címnek csupán egy-egy szavát ismeri, de más adatot nem tud.

A Permuterm Subject Index az index-gyűjtemény nélkülözhetetlen része. Nélküle a Science Citation Index rendszere kétségtelenül hiányos. Ha a Source Index-et, mely a szerzők nevei szerinti közelítés eszköze, a gyűjtemény alapvető részének tekintjük, akkor ahhoz az egyik oldalról a Citation Index csatlakozik, hogy az irodalmi utalások láncolatának igénybevételével történő közelítést biztosítsa, a másik oldalról pedig a Permuterm Subject Index, mely viszont a címszavak felhasználásával folyó közelítést teszi lehetővé.

A két utóbbi rész ugyanis csak a Source Indexszel együtt használható. A Source Index szolgáltatja az azokban szereplő forrásjelzetekhez a vonatkozó forrás-dokumentumok címadatait.

A Permuterm Subject Index az egész index-gyűjtemény számára új alkalmazási területet nyitott, s ezzel a gyűjtemény értékét erősen fokozta.

A Permuterm Subject Index azonban a Source Index-szel együtt - a Citation Indexet elhagyva - címszavak szerinti keresésre önálló indexmódként is használható.

A Permuterm Subject Index-ben a cikkek a szerzők által a címekben használt szavakkal szerepelnek. Az index hűen, sőt több mint hűen: mechanikusan követi a szerzők szóhasználatát.

Ebből egyaránt adódnak előnyös és előnytelen következmények.

Előnye az indexnek, hogy

nem befolyásolja a szavak értelmének módosulása;

az új fogalmak szóképei, az új szakkifejezések azonnal helyet kapnak benne.

Viszont hátrányos körülményt képez az, hogy

a szinonimák nem iktathatók ki, s ezért ugyanaz a fogalom az indexben egyidejűleg több szóképpel szerepelhet;

a homonimák keveredését nem lehet megakadályozni. Valamennyi dokumentum megjelenik a címében előforduló szóképek alatt, tekintet nélkül arra, hogy a szóképek az egyes dokumentumok címeiben mi a jelentése.

Az index jövőbeli évfolyamaiban a tárgyi keresés hatékonyságát fokozni óhajtják. Ennek érdekében a szerkesztőség tervbe vette a vezérszavak közötti keresztutalások rendszerének kiépítését.

AZ INDEX SZERKEZETE

Az 1. ábrán az index egy lapjának részletét mutatjuk be; ezen az index szerkezete jól szemlélhető.

Az index kétfokozatu. Indextételei két-két tárgyi fogalomból tevődnek össze: az elsődleges és a társfogalomból /primary term, co-term/. Az elsődleges fogalmakat vezérszóként vetítik ki, s alattuk következnek oszlopba rendezetten a társfogalmak. Az elsődleges fogalmak adják az indexben való keresés főszempontját. A közelítés pontosabbá tételét az elsődleges fogalmak alatt közölt több-kevesebb társfogalom szolgáltatja.

A dokumentumok azonosítása két adattal történik: az első szerző nevével és a dokumentum azonossági számával.

A teljes indextétel tehát négy adatfelelősségből tevődik össze, amelyek közül a három utolsó oszlopai alakítják az indexoldalak hasábjainak jellegzetes háromoszlopos strukturáját.

Az indextételek helyes egymásutánját három betűrendes és egy szám szerinti sorrend biztosítja. Az index egészének vezérlőrendjét az elsődleges fogalmak betűrendje szolgáltatja. Ezen belül a társfogalmak betűrendje érvényesül; amelyen belül pedig a szerzők nevei

következnek hasonló rendben. Ugyanazon szerző több publikációja esetén az azonossági számok emelkedő egymásutánban sorakoznak egymás alatt.

Egyes társfogalmaktól a hozzájuk tartozó szerző nevére nyíll mutat. Ezzel azt tüntetik fel, hogy a szóbanforgó dokumentum az elsődleges fogalom vonatkozásában az indexbe felvett valamennyi irodalmi forráshely adatait tartalmazza. Ha tehát valaki az elsődleges fogalomra vonatkozó összes tanulmány bibliográfiai adatait óhajtja, a nyíllal jelölt dokumentumban megtalálhatja azokat.

Az első kötet elejére az indexelt folyóiratokról ötféle, szintén computerrel készített és kiírt jegyzéket helyeztek, amelyek jó segédeszközök az index használatához. A jegyzékek a következőket tartalmazzák:

- a folyóiratnevek rövidítéseit, a teljes nevekekkel;
- a folyóiratok neveit, rövidítéseikkel;
- az index jelen évfolyamában feldolgozott folyóiratszámokat, azonosító jelzeteikkel;
- a folyóiratokat országonkénti csoportosításban;
- a folyóiratokat szakonkénti csoportosításban. Ebben a jegyzékben egyes folyóiratok több csoportban is szerepelnek, ha tartalmuk ezt indokoltá teszi.

A Permuterm Subject Index hat kötetében 1573 folyóirat összesen 12 444 füzetét dolgozták fel. A műben 287 031 közleményt indexeltek. Az index oldalai hat hasábra rendezetten tartalmazzák az indextételeket. A hat kötet együttesen 37 380 hasábot ölel fel. Mivel a hasárok kétszáz sorosak, az indextételek összmennyisége meghaladja a hét milliót. Dokumentumonkénti átlagos mennyiségük harminc.

A Permuterm Subject Index az eddig publikált tárgyi indexek között terjedelemben a leghatalmasabb, nyomdatechnikailag a legkompaktabb, adatvolumene a legnagyobb.

Az elsődleges fogalmak maximális terjedelme 18 jegy lehet. Ha a fogalom ennél több betűből áll, akkor a 18. pozícióra pont kerül. Ez mutatja, hogy a fogalom többi jegyeit elhagyta a gép.

A társfogalmak a szöveghasámban az elsődleges fogalomhoz képest két pozícióval beljebb kezdődnek. Terjedelmüket 11 pozícióban korlátozták. Az ennél hosszabb társfogalmak utolsó jegyeinek elhagyását feltüntető pont a 11. pozícióra kerül. Ha a társfogalmak a maximális terjedelemtől rövidebbek, szaggatott vonal tölti ki a sornak a 11. pozíciótól terjedő részét.

A szerzők nevét felsoroló oszlop szintén 11 pozíció szélességű. Tartalmazza az első szerző vezetéknevét 8 betű terjedelemtől, valamint keresztnévét közül annyinak első betűit egymás mellé írva, amennyi a rendelkezésre álló pozíciókon elhelyezhető. Ha a vezetéknev

8 betűnél hosszabb, további betűi elmaradnak, s ennek jeléül a 9. pozícióra pont kerül.

Az anonim közlemények esetében a szerző helyett a folyóirat nevéből közölnek 11 pozíciónyi részletet, illetve a folyóirat nevének rövidítését maximálisan 11 pozíciónyi terjedelemben. A rövidítések feloldását az első kötet elején közölt jegyzék adja.

Az utolsó 6 pozíciós oszlop a dokumentumok azonossági számát tartalmazza. A hat pozícióból az első öt pozíción található

a szerzős dokumentumok esetében a vonatkozó folyóiratfüzetnek un. ISI-száma, azaz a folyóiratfüzetnek az Institute for Scientific Information-ban a tárgyévra adott gyarapodási száma, mely öt számjegyből áll;

az anonim dokumentumoknál viszont az első pozíción P betű áll, jelölve, hogy utána lapszám következik: a dokumentum első oldalának a folyóiratévfolyamon belüli lapszáma. A lapszám maximálisan négy jegy terjedelmű lehet.

A 6. pozíción egyes dokumentumok mellett egyetlen betűből álló kódjel szerepel. A betűkód a különleges dokumentumtipusok megjelölésére szolgál.

A következő hét kódot alkalmazzák:

D = discussions, conference items
/vitacikkek, konferenciák előadásai/

E = editorials
/szerkesztői megjegyzések, üzenetek/

L = letters, communications
/a szerkesztőséghez intézett levelek, rövid közlemények/

M = proceedings from meetings
/értekezletek, kongresszusok kiadványai/

Q = bibliography
/bibliográfia/

R = reviews
/szemlék/

A tanulmányok, kutatási jelentések /reports/, technikai leírások /technical papers/ mellett betűkód nem szerepel.

AZ INDEX ELŐÁLLÍTÁSA

A tárgyi indexeknek a dokumentumok címeiből elektronikus adatfeldolgozó berendezéssel történő automatikus előállítására eddig már

több eljárást dolgoztak ki, s ezek nyomán gépi uton különböző típusú indexeket állítottak elő. Ezekhez a Permuterm Indexszel egy további járult.

A Permuterm-eljárás a dokumentumok címeinek a téma meghatározása szempontjából jelentősnek minősülő szavait felhasználva állítja elő az indexet.

Az angol nyelv alaktanának egyszerűsége lehetővé teszi, hogy a címekben szereplő jelentős szavak eredeti formájukban, minden változtatás nélkül vezérszavakként alkalmazhatók. Ez szolgáltatja az alapot a címek szókészletének computerrel végzett automatikus feldolgozásához.

Más nyelvek esetében a szavak toldalékai /ragok, képzők, igekötők stb./ miatt, valamint a töváltozások nagy száma folytán a Permuterm-eljárás annyira nehézkessé válik, hogy gyakorlatilag elveszti alkalmazhatóságát.

A Permuterm-eljárás a címek szókészletében négyféle szótípust különböztet meg:

teljesen irreleváns szavakat /full stop words/, amelyek kizárólag a szöveg nyelvtani vagy stílusi jellegessége szempontjából szükségesek. Összesen 115 szót minősítettek ilyennek. Az ezekből összeállított kis szótár a 2. ábrán látható;

Full Stop

A	CONCERNING	MORE	THEREFROM
ABOUT	DISCUSSION	MY	THESE
ABOVE	DO	NEAR	THEY
ACROSS	DOWN	NEARLY	THIS
AFTER	DR	NEXT	THOSE
AGAINST	DUE	OF	THROUGH
ALL	DURING	OFF	THROUGHOUT
ALMOST	EITHER	ON	TO
ALONG	FEW	ONLY	TOGETHER
ALSO	FOR	ONTO	TOWARD
AMONG	FROM	OR	TOWARDS
AN	FURTHER	OTHER	UNDER
AND	GIVE	OUR	UP
ANOTHER	GIVEN	OUT	UPON
ANY	GIVING	OUTSIDE	UPTAKE
ARE	HAVE	OVER	UPWARD
AROUND	HAVING	PER	VERY
AS	HIS	REGARDING	WHAT
AT	HOW	SAME	WHATS
BE	IN	SEEN	WHEN
BEFORE	INSIDE	SEVERAL	WHERE
BEHIND	INSTEAD	SHOWN	WHICH
BELOW	INTO	SINCE	WHILE
BETWEEN	IS	SOME	WHO
BEYOND	IT	THAN	WHOM
BOTH	ITS	THAT	WITH
BUT	LITTLE	THEIR	WITHIN
BY	MADE	THEM	WITHOUT
CERTAIN	MANY	THERE	

2. ábra

Az irreleváns szavak szótára

félig irreleváns szavakat /semi-stop words/, amelyek az index-tételekben elsődleges fogalmak gyanánt nem szerepelhetnek, de társ-fogalmak lehetnek. Jegyzékükbe 227 szót vettek fel. Egy részüket a 3. ábra közli;

Semi-Stop

ACTIVITY	CAUSING	CONTENTS	EFFECTIVENESS
ACTIVITIES	CENT	CONTRIBUTION	EFFECTS
ACTION	CERTAIN	CONTRIBUTIONS	ESTIMATE
ACTIONS	CHANGE	CONTROL	ESTIMATED
ACUTE	CHANGES	CONTROLLED	ESTIMATES
ADDENDUM	CHARACTERISATION	CONTROLLING	ESTIMATING
ADVANCES	CHARACTERISTIC	CONTROLS	ESTIMATION
AFFECTING	CHARACTERISTICS	CORRECTION	ESTIMATIONS
ALTERATION	CHARACTERIZATION	CORRECTIONS	EVALUATION
ALTERATIONS	CHARACTERIZATIONS	CORRELATION	EVALUATIONS
ANALYSES	CHARACTERIZED	CORRELATIONS	EVIDENCE
ANALYSIS	CHARACTERIZING	CORRESPONDENCE	EXAMINATION
ANALYTICAL	CLASS	COURSE	EXAMINATIONS
ANNUAL	CLASSES	D	EXPERIMENT
ANON	COMMENT	DEGREES	EXPERIMENTAL
APPLICATION	COMMENTS	DEMONSTRATED	EXPERIMENTALLY
APPLICATIONS	COMPARATIVE	DEMONSTRATION	EXPERIMENTS
APPLIED	COMPARISON	DETERMINATION	FACTOR
APPROACH	COMPARISONS	DETERMINATIONS	FACTORS
APPROACHES	COMPOSITION	DETERMINE	FINDINGS
ASPECT	COMPOSITIONS	DETERMINED	FOLLOWED
ASPECTS	COMPOUNDS	DETERMINING	FOLLOWING
ASSESS	CONCENTRATION	DEVELOPING	FORMATION
ASSESSING	CONCENTRATIONS	DEVELOPMENT	FORMATIONS
ASSESSMENT	CONCEPT	DEVELOPMENTS	FUNCTION
ASSOCIATED	CONCEPTS	DIFFERENCES	FUNCTIONS
BASED	CONCERNING	DIFFERENT	G
BASIS	CONDITION	DIRECT	GENERALIZATION
CALCULATION	CONITIONS	E	GENERALIZATIONS
CALCULATIONS	CONFERENCE	EARLY	IMPLICATION
CASE	CONGRESS	EFFECT	IMPLICATIONS
CASES	CONSIDERATION	EFFECTING	IMPORTANCE
CAUSE	CONSIDERATIONS	EFFECTIVE	INCIDENCE
CAUSED	CONTAINING	EFFECTIVELY	INCIDENCES
CAUSES	CONTENT		INDIRECT

3. ábra

A félig irreleváns szavak szótárának részlete

irreleváns szópárokat /paired word stop/; 33 szópárt gyűjtöttek össze, amelyek bár a címek formai teljességéhez hozzátartoznak, de a címek tárgyi tartalma szempontjából teljességgel jelentéktelenek. Ha a címekben ezek közül valamelyik előfordul, a computer nem veszi figyelembe.

Ilyen szópárok:

brief - communications	/rövid - közlemények/
current - status	/jelenlegi - helyzet/
improved - method	/javított - módszer/
papers - selected	/válogatott - tanulmányok/

particular - reference	/különös - tekintettel/
preliminary - report	/előzetes - jelentés/

stb.;

releváns szavakat. A feldolgozott címnek minden olyan szava, mely az előbb említett három csoport egyikében sem található, relevánsnak minősül. A releváns szavak az indexben mind elsődleges, mind társfogalmak gyanánt szerepelnek.

A computer minden feldolgozásra kerülő dokumentumcím valamennyi szavát összeveti az előbbieken említett három szógyűjtemény szókészleteivel. Azokat a szavakat, amelyek az első és harmadik szógyűjteményben előfordulnak, a további műveletek során mellőzi. A második szógyűjteményben található szavakat pedig csak társfogalomként alkalmazza. A három szógyűjtemény egyikében sem található szavakat viszont mind elsődleges, mind társfogalmak gyanánt kezeli.

A szavak típusainak meghatározása után permutálással szópárokat állít elő a gép. A szópárok első szava az indexben elsődleges fogalom lesz, második szava társfogalom. Majd a szópárok mindegyikéhez csatolja a dokumentum azonossági jelét. Az így elkészített indextételeket a már említett rendezési elvek szerint szigorú sorrendbe osztja.

A rendezés befejeztével a computer gyorsnyomtató készülékével kiírja az index szövegét.

A feldolgozás menetének illusztrálására tekintsük az alábbi példát:

A dokumentum címe: Controlling variations in mechanical properties of heat resistant alloys during forging /Hőálló ötvözetek mechanikai tulajdonságaiban kovácsolás folyamán bekövetkező változások ellenőrzése/.

Szerzők: DERIDDER, A.J. és KOCH, R.W.

Forrás: Metals Engineering Quarterly, 5.k. 3.sz. 1965. kezdőold. 61.

Azonossági jelle: Deridder AJ 71477

A dokumentumcím szókészlete így oszlik meg:

irreleváns szavak: during, in, off

félíg irreleváns szavak: controlling, properties, variations

releváns szavak: alloys, forging, heat, mechanical, resistant.

A dokumentumcíméből permutálással képzett elsődleges fogalom - társfogalom szópárokat a 4. ábra tartalmazza. A dokumentum ezzel a 35 szópárral szerepel a Permuterm Subject Indexben.

<u>alloys</u>	<u>forging</u>	<u>heat</u>
controlling	alloys	alloys
forging	controlling	controlling
heat	heat	forging
mechanical	mechanical	mechanical
properties	properties	properties
resistant	resistant	resistant
variations	variations	variations
	<u>mechanical</u>	<u>resistant</u>
	alloys	alloys
	controlling	controlling
	forging	forging
	heat	heat
	properties	mechanical
	resistant	properties
	variations	variations

4.ábra

A Controlling variations in mechanical properties of heat resistant alloys during forging cím elsődleges fogalmi társfogalmaikkal

AZ INDEX KRITIKAI VIZSGÁLATA

A Permuterm Subject Index több különböző szempontból vethető kritikai vizsgálat alá. Interdiszciplináris jellegével, valamint az a kérdéssel, hogy a különféle ismeretágak szempontjából milyen foku a használhatósága - illetve, ami ennek alapját képezi, hogy a feldolgozott folyóiratállomány milyen arányban oszlik meg az egyes szakterületek között - nem foglalkozunk. Ezeket a kérdéseket a Citation Index vonatkozásában a folyóirat egyik tanulmánya már tárgyalta /2/.

A Permuterm Subject Indexet kizárólag mint a tárgyi keresés eszközt vizsgáljuk.

E nézőpontból két kérdés vethető fel:

megfelel-e az index rendeltetésének?
más index-struktúra alkalmasabb volna-e?

Az első kérdésre határozott igennel kell felelni. Kétségtelen, hogy hiánytalanul adja azt, amit kétfokozatu szerkezetével nyújthat. Kezelése egyszerű. A feldolgozott címanyagból mindenkor teljes eredményt szolgáltat, mivel minden szópár, melynek nyomán keresés történhet, indextelként megtalálható benne.

Az index szövege gyakorlatilag hibátlanak tekinthető. Sokszor használtuk az indexet, s csupán egyetlen hibát találtunk. Es természetesen nem jelenti azt, hogy több nem lenne, de ha akad is, mennyiségük egészen elenyésző, s az index használatát kedvezőtlenül nem befolyásolja.

Egyetlen komoly kifogásként az indextételekben megnyilvánuló redundancia hozható fel. A computer a címnek releváns és félig irreleváns szavaiból az index elsődleges és társfogalmait tartalmazó szópárokat nemcsak automatikusan, hanem mechanikusan is képezi. Programja szerinti eljárással kialakítja az összes szópárokat, tekintet nélkül arra, hogy a szópárokat alkotó szavak ésszerűen egymáshoz kapcsolhatók-e vagy sem. Így számos olyan szópár létesül, amelynek a szóbanforgó dokumentumcím vonatkozásában indextételenkénti alkalmazása nem indokolt. Sőt, ha inkompatibilis fogalmak kerülnek egymás mellé, értelmetlen szópárok kerülnek az indexbe.

Tekintsük például az alábbi dokumentumcímét:

A computer controlled information retrieval system for an individual investigator /Computerrel vezérelt információkereső rendszer az egyedi kutató részére/. Azonosítási jele: COOK RW 87017.

A címben szereplő "individual" jelző az "investigator" főnévre vonatkozik. Tehát az "individual" elsődleges fogalom alatt társfogalomként kizárólag az "investigator" szerepeltetése indokolt. Viszont a mechanikus eljárás folyományaként az "individual" szó, mint elsődleges fogalom még az alábbi társfogalmakkal is képez indextételeket:

individual
computer
controlled
information
retrieval
system

Ezek a szópárok azonban a dokumentumcím vonatkozásában érdektelenek, sőt nem egyszer értelmetlenek is.

A felesleges szópárok mennyisége címenként változó, de általában tetemes. Ennek következtében az indexben számos olyan indextétel jelenik meg, amely a dokumentum témája szempontjából hasznos információt nem szolgáltat. Ezek az indextételek nemcsak indokolatlanul duzzasztják az indexet, hanem egyenesen károsak, mert hamis információikkal félrevezetik a kutatót. Arról, hogy az ilyen indextételek információi helytelenek, az informálódó legtöbbször további fáradozás árán tud csak meggyőződni; éspedig jobbik esetben akkor, ha a vonatkozó dokumentumcímeket a Source Indexből kikereszte, rosszabbik esetben pedig, miután magukat a dokumentumokat átnézte.

Az indextételek ártalmas redundanciáját megszüntetni csak emberi közreműködéssel lehet. A géppel kialakított szópárok relevanciáját ugyanis a dokumentumcím vonatkozásában mérlegelni kell, amire a computer nem képes. Az anyag hatalmas volumene miatt viszont a szakértői revízió megvalósítása útközik gyakorlatilag áthidalhatatlan akadályokba.

Az utóbbi tény, valamint az index teljesen automatizált előállításából fakadó számos jelentős előny arra kényszerítette a szerkesztőket, hogy lemondjon az intellektuális szűrésről, illetve az indexnek ilyen úton elérhető, a nagyságrendileg számottevő érdemi javításáról. Ezért szerepel az indexben minden géppel létrehozott szó pár, akár hasznos, akár felesleges, vagy éppenséggel ártalmas.

A Permuterm Subject Indexszel, mint a tárgyi keresés eszközevel kapcsolatos másik kérdés annak tisztázására irányul, hogy módosított szerkezettel jobban megfelelne-e rendeltetésének.

Az index szerkezetén kétféle változtatás hajtható végre.

Az egyik változtatás abban nyilvánulhatna meg, hogy az index-tételekben az azonossági jel helyett a dokumentumcimet szerepeltetnék. Ez az indexben való keresést kétségtelenül sokkal hatékonyabbá tenné, mert az indextételek kimerítő tárgyi információt szolgáltatnának. A Source Index igénybevétele vagy teljesen feleslegessé válhatna, vagy csupán ama néhány dokumentum bibliográfiai adatainak megtekintésére korlátozódhatna, amely dokumentumok címeik nyomán az információt kereső számára felhasználhatónak minősülnek. Ilyen körülmények között a Permuterm Subject Indexet az indexsorozat többi részeitől független, önállóan használható keresési eszközzé lehetne tenni.

Ez az indexstruktúra a jelenleginél feltétlenül fejlettebb volna. Computerrel való automatikus előállítása sem ütköznék nehézségbe. Megvalósítása esetén azonban a Permuterm Subject Index jelenlegi volumenének legalább tíz-tizenötszörösére emelkednék. Ekkora ugyanis az átlagos arány a dokumentumcímek és az azonossági jelek terjedelme között.

Az index szerkezetének ilyen irányú megváltoztatása tehát valószínűsíthető, bár elvileg nemcsak megoldható, hanem egyenesen előnyös.

A szerkesztésmódosítás másik területét az index fokozatai szolgálhatnák, mégpedig alternative két irányban: a meglévő fokozatok csökkentésével, illetve növelésével. Az index ugyanis jelenlegi két fokozata helyett vagy csupán egyetlen, vagy kettőnél több fokozattal is előállítható lenne. Vizsgáljuk meg mindkét lehetőséget.

Az egyfokozatu megoldásban az indextételek egyetlen tárgyszót tartalmaznak. A tárgyszó alatt sorakoznak azoknak a dokumentumoknak azonossági jelei, vagy - ha az index komplett információt akar szolgáltatni - címei, amely dokumentumok címeinek a tárgyszó egyik releváns szava. Ennél a megoldásnál a dokumentumcímek annyi indextétellel szerepelnek az indexben, ahány releváns szó található bennük.

Az egyfokozatu szerkesztés esetén, különösen akkor, ha az index-tételekben a dokumentumra csak az azonossági jel utalna, az index volumene nagymértékben csökkenne. De elfogadható keretek között maradna még akkor is, ha az indextételekben a teljes dokumentumcimet

közölné. Ilyen szempontból tehát az egyfokozatu struktura alkalmas megoldást képezne.

Égészen másképpen alakulna azonban a helyzet az index használhatósága szempontjából. Az egyetlen fogalomra korlátozódó vezérszavak alatt gyakran több tiz, de nem egyszer több száz irodalmi hivatkozás gyűlné össze. Információkeresés esetén ezek mindegyikének irodalmi utalását át kellene nézni, mert csak így volna megállapítható, hogy a vonatkozó dokumentum a keresett információ szempontjából kielégítő-e vagy sem. Ez még akkor is nehézkessé tenné az informálódást, ha az indextételek tartalmaznák a dokumentumcimet. Abban az esetben azonban, ha az azonossági jelek nyomán a Source Indexből kellene minden dokumentumcimet kikeresni, a hivatkozások nagy száma miatt az index gyakorlatilag használhatatlanná válnék.

Vizsgáljuk meg, mit jelentene az ellenkező irányu fokozatváltoztatás.

Az index fogalmainak kettőről háromra, négyre, való emelése esetén a dokumentumcimből kialakított szőláncok képeznék az indextételeket. Kétségtelen, hogy ebben az esetben a dokumentumcimekben szereplő témák az indextételekben jól kifejezésre jutnának; mégpedig minél hosszabb lenne a szőlánc, azaz minél több fokozatu lenne az index, annál jobban. Mivel a szőláncok betűrendezése azok egész hosszára kiterjed, ez már önmagában is tagolt indexstruktúrát eredményezne. Lépcsőzetes, illetve oszlopos felépítés esetén az áttekinthetőség még fokozódna. További előnyt pedig az képezne, hogy a három-négy tárgyszót felölelő indextételek legtöbbször önmagukban elegendők lennének annak elbírálására, hogy a vonatkozó dokumentum a megválaszolandó kérdés szempontjából felhasználhatónak minősül-e vagy sem. Így minimumra csökkenne azoknak az eseteknek száma, amikor a dokumentum témája felőli részletesebb informálódás érdekében a Source Indexhez kell folyamodni.

Az előnyök ellenében azonban igen kellemetlen hátrányok jelentkeznek. A fokozatok számának emelésével hatványozottan emelkednek az egyes dokumentumcimekből képezhető indextételek száma, ennek viszont egyik következménye az index már amúgy is túlméretezett terjedelmének erőteljes bővítése lenne. Becslésünk szerint elérné a jelenlegi terjedelem három-négyszeresét. A másik, sokkal súlyosabb kihatásu következmény, hogy a szőláncok mechanikus kialakításának folyományaként sok, téves információt szolgáltató vagy irracionális szókapcsolatot tartalmazó indextétel keletkeznék. Vizsgálataink szerint ezek mennyisége olyan magas lenne, hogy gyakorlatilag illuzórikussá tenné az információkeresés eredményességét.

Következésként tehát azt a megállapítást rögzíthetjük, hogy az indexfokozatok számának emelése az index használatát egyrészt hatékonyabbá tenné, másrészt károsan befolyásolná. Ha úgy tekintjük, hogy e két ellentétes hatás, amelyet pontosan felmérni amúgy sem lehet, egymást kiegyenlíti, akkor is marad még egy negatívum, az index túl nagy volumene.

Második kérdésünkkel kapcsolatban tehát vizsgálataink azt a meg-

állapítást szolgáltatják, hogy a Permuterm Subject Index szerkezetének bármilyen irányú módosítása az elérhető pozitívumok mellett azoknál sokkal nyomósabb negatívumokat eredményezne.

Kellő megfontolással végiggondolva azokat a tényezőket, amelyek a Permuterm Subject Index szerkezetének kialakításánál dominánsak voltak, be kell látnunk, hogy azt ésszerűen kiképezni csak meglévő formájában lehetett.

S^SS

I R O D A L O M

1. Hazai irodalma:

BASSOLA Z.: A Science Citation Index. = Könyvtári Figyelő, 11.k. 5-6.sz. 1965. p.382-392.

POLZOVICS I.: Kísérlet indexművek szerkesztésének mechanikus megoldására. A Science Citation Index. = Tudományos és Műszaki Tájékoztatás, 14.k. 12.sz. 1967. p.819-841.

Külföldi irodalmát POLZOVICS I. tanulmányához csatolt bibliográfia közli.

2. POLZOVICS I.: i.m.

W^WW

OROSZ, G.: The Permuterm Subject Index

The paper reviews the Science Citation Index and points out especially the role and significance of the Permuterm Subject Index being a tool for literature searching under title entries.

The entries consist of four components: two terms /concepts/, the author's name and the identifying symbol /indicator of identity/ of the document. The first term is the primary term, the second is the co-term. The primary term is used as a weighted keyword followed by the corresponding co-terms, while the co-term is followed by the other two components.

The methods of processing the index are the following: entries are automatically generated from the title by computer. Three machine vocabularies are used consisting 1. of full stop words /irrelevant words/; 2. of semi-stop words /semi-irrelevant words/; and 3. of paired word stops /irrelevant paired words/. The computer, based upon the vocabularies, selects the relevant and semi-stop words from the title entries. The remaining relevant and semi-stop words in the title are assigned as co-terms to each of the primary relevant words, thus forming pairs of primary term - co-term. The computer adds the author's name and the identification sign of the document to these pairs. The complete entries, finally, are organized into the stipulated order.

The critical examination of the Permuterm Index shows that the information content of the titles is being revealed by a maximum amount of entries. Therefore the index is complying with its intended purpose. Although the entries taken from the title contain a relatively small amount of information, there is a possibility of tracing practically all relevant words in the title. However, there is a serious drawback: the number of superfluous entries is considerable, which means that the computer generates all the primary term - co-term pairs from the title, irrespective of their relevance from the viewpoint of the content. The redundancy of title entries could only be eliminated by human effort: this however, - owing to the necessary quantity of mental work - is far too circumstantial and costly and therefore practically not feasible.

The paper raises the question, whether the index could be rendered more usable by changing its structure.

One possible modification would be if the entries contained the full titles of the document instead of only the identification sign. This solution, however - even if it might seem useful theoretically - seems unrealistic due to the considerable increase in the size of the index.

Another possible change would be: to decrease or increase the number of title entries. In the first case the entry would include only one subject term. This structure would be obviously inferior to the existing one, since references under individual entries /concepts/ would accumulate to an extent that searching would be practically impossible. An increase in the number of title entries, in turn, would only result in a considerable growth of the number of entries; consequently, both the size of the index and the number of superfluous entries would grow considerably and searching would become very difficult.

If we take into consideration the predominant aspects when determining the structure of the Permuterm Subject Index, it must be admitted that the present structure seems to provide the best possible solution.

∴∴∴

ОРОС, Г.: Предметный указатель Пермутерм

В вводной части статьи представляются система комплекса Science Citation Index /Указатель цитированной литературы по точным, естественным и прикладным наукам/, кроме этого роль и значение размещенного в ней Permuterm Subject Index, средства поиска по заглавным словам.

В дальнейшем автор занимается построением указателя. Каждая рубрика указателя /index entry / состоит из четырех компонентов: из двух предметных слов, имени автора и опознавательного кода. Первое из предметных слов является первичным словом /primary term /, а второе сопряженным словом /co-term/. Первичное слово выделено, как заголовок, под ним получают место принадлежащие к нему сопряженные слова. Рядом с сопряженным словом размещаются два прочих компонента. В следующей части статьи представляются технические характеристики указателя и данные его объема.

В дальнейшем автор статьи занимается методом составления указателя. Вычислительная машина автоматически выделяет рубрики указателя из заголовка. К этой работе применяются три машинных словаря: словарь незначащих слов /full stop words/, словарь полу-незначащих слов /semi stop words/, словарь незначащих пар слов /paired word stop/. При помощи этих вычислительная машина избирает значение и полу-значение слова из заголовка. После этого к каждому значащему слову, как первичному слову приписывается в качестве сопряженных все остальные значения и полу-значения слова заголовка: таким образом образуются пары слов, состоящие из первичного и сопряженного слов. Вычислительная машина приписывает к парам слов имя автора и опознавательный код. Наконец вычислительная машина по предписанию упорядочивает уже готовые рубрики.

Далее автор занимается критическим рассмотрением указателя. Информационное содержание заголовка выявляется максимальным числом рубрик указателя. Это обеспечивает соответствие указателя своему назначению. Рубрики указателя содержат сравнительно мало информации из заголовка. Зато имеется возможность приближения со всех существенных слов заголовка. Единственным серьезным возражением является наличие большого числа лишних рубрик указателя. Вычислительная машина выделяет из заголовка все пары слов, состоящие из первичного и сопряженного слов, в независимости от их значимости с точки зрения передачи содержания заголовков. Пресечь такую избыточность рубрик указателя можно было только посредством живого человеческого труда. Это, однако, ввиду большого объема требует слишком много времени и затрат, таким образом практически не осуществимо.

Автор рассматривает и другую проблему: улучшился ли указатель при другой структуре.

Одним изменением могло бы быть помещение названия документа на место опознавательного кода. Такое решение, хотя в принципе и было бы полезным, нереально, так как привело бы к значительному увеличению объема.

Можно представлять и другое изменение: уменьшение или увеличение числа предметных слов. В первом случае рубрика указателя заключала бы в себе только одно предметное понятие. Очевидно такая структура была бы хуже настоящей, так как под рубриками указателя собралось бы такое большое число ссылок, что практически было бы невозможно вести поиск по указателю. При увеличении числа заглавных слов рубрик указателя, однако, сильно увеличилось бы число рубрик, могущих быть образованными из каждого заголовка. От этого с одной стороны увеличился бы и без того большой объем указателя, а с другой стороны же - возросло бы число лишних рубрик настолько, что поиск стал бы иллюзорным.

Приняв во внимание, таким образом, все те соображения, которые доминировали при выборе структуры Permuterm Subject Index необходимо признать, что этот указатель можно было построить исключительно в настоящей его форме.

.. 0 ..

OROSZ, G.: Der Permuterm Subject Index

In der Einführung wird das System des Science Citation Index besprochen und die Rolle des Permuterm Subject Indexes als Mittel für die Literaturrecherche auf Grund von Stichwörtern dargestellt.

Die Untersuchung der Struktur des Indexes zeigt, dass die Indextitel /entry/ aus vier Bestandteilen bestehen: zwei Sachbegriffe, der Name des Autors und das Identitätszeichen des Dokuments. Von den Sachbegriffen ist der erste der primäre Begriff /primary term/, der zweite der Mitbegriff /co-term/. Der primäre Begriff ist als Ordnungswort hervorgehoben und darunter folgen die dazugehörigen Mitbegriffe. Neben dem Mitbegriff befinden sich dann die weiteren zwei Bestandteile.

Der Computer bestimmt die Indextitel automatisch aus dem Titel.

Dazu werden drei Wortverzeichnisse verwendet: das Verzeichnis 1. der irrelevanten Wörter /full stop words/, 2. der semi-irrelevanten Wörter /semi-stop words/ und 3. das der irrelevanten Wortpaare /paired word stop/. Mit Hilfe der Verzeichnisse wählt der Computer aus dem Titel die irrelevanten und semi-irrelevanten Wörter aus. Die ausgewählten relevanten Wörter werden als primäre Begriffe geordnet, die übrigen relevanten, sowie irrelevanten Wörter werden als Mitbegriffe zugefügt. Auf diese Weise werden Wortpaare primärer Begriff - Mitbegriff gebildet. Den Wortpaaren ordnet der Computer den Namen des Autors und das Identitätszeichen zu. Die fertigen Indextitel werden vom Computer in der vorgeschriebenen Reihenfolge geordnet. Der Informationsgehalt der Titel wird durch eine maximale Zahl von Indextiteln erschlossen. Dadurch wird gesichert, dass der Index seinem Zweck entspricht. Die Indextitel enthalten jedoch verhältnismässig wenig Information aus dem Titel. Es besteht aber die Möglichkeit die im Dokument enthaltene Information mit allen relevanten Wörtern des Titels aufzuschlüsseln. Die einzige ernsthafte Einwendung wäre, dass im Index zahlreiche überflüssige Indextitel zu finden sind. Der Computer bildet nämlich sämtliche primärer Begriff - Mitbegriff Wortpaare ohne Rücksicht darauf, ob jene vom inhaltlichen Gesichtspunkt wichtig oder wertlos sind. Diese Redundanz der Indextitel könnte nur durch menschliche Tätigkeit abgeschaffen werden. Dies ist aber wegen der notwendigen grossen geistigen Arbeit zu kostspielig, also praktisch undurchführbar.

Der Autor untersucht ob der Index mit abgeänderter Struktur besser geeignet wäre.

Eine Änderungsmöglichkeit könnte darin bestehen, dass die Indextitel statt des Identitätszeichens den Titel des Dokuments voll angeben würde. Diese Lösung - obwohl prinzipiell nützlich - ist wegen der bedeutenden Zunahme des Umfanges trotzdem unrealistisch.

Eine andere Änderungsmöglichkeit könnte durch die Verminderung oder Erhöhung der Zahl der Stichwörter durchgeführt werden. Im ersten Fall würde der Indextitel nur einen einzigen Sachbegriff enthalten; das ergebe eine offenbar schlechtere Struktur als die gegenwärtige, da sich unter den Indextiteln so viele gleiche Stichwörter anhäufen würden, dass dadurch das Suchen im Index praktisch unmöglich wäre. Infolge der Erhöhung der Anzahl der Stichwörter würde dagegen die Anzahl der aus den Titeln bildbaren Indextitel bedeutend zunehmen. Demzufolge würde sich einerseits das sowieso grosse Ausmass des Indexes erhöhen, andererseits aber die Anzahl der überflüssigen Indextitel dermassen vergrössern, dass die Informationsrecherche dadurch illusorisch würde.

Es muss auf Grund der Gesichtspunkte, welche bei der Festlegung der gegenwärtigen Struktur des Permuterm Subject Indexes vorherrschten, festgestellt werden, dass dieser nur in der derzeit bestehenden Form zusammengestellt werden konnte.
