025.4:681.3

Classificação automática: revisão da literatura

GLAUCIA HELENA BARBOSA PEREIRA DE SOUSA *

Revisão da literatura na área da classificação automática: trabalhos mais significativos, seus pesquisadores, os diferentes métodos experimentados.

A classificação automática de documentos ainda representa em nossos dias um desafio aos estudiosos do assunto.

Se esta tarefa pudesse ser realizada num alto nível de precisão através do uso da máquina (tarefa esta, cuja realização vem sendo pesquisada desde os anos cinqüenta por diversos cientistas), um dos mais graves problemas de nosso século estaria solucionado, isto é, o atraso no processamento do documento, que conseqüentemente provoca a demora na entrega do mesmo ao usuário interessado.

O problema da classificação automática de documentos é uma parte do problema mais amplo na análise automática do conteúdo (uma tarefa complexa envolvendo a análise do significado). Classificação foi

* Bibliotecária do Centro de Informações Tecnológicas do Instituto Nacional de Tecnologia, Rio de Janeiro, GB.

definida por Borko (1963)[7] como "a determinação do conteúdo do assunto".

Um levantamento da literatura nesta área levou-nos às seguintes conclusões: o trabalho pioneiro foi o de Luhn (1957)[25] onde ele demonstrou que uma análise estatística das palavras nos documentos proporcionaria algumas indicações quanto ao seu conteúdo. Ele também citou o processo de se comunicar conceitos através de palavras, e sugeriu que, "quanto mais duas representações concordassem em dados elementos e sua distribuição, maior seria a probabilidade delas representarem informação semelhante".

Maron (1961) [26] também apresentou um ponto de vista estatístico e propôs estabelecer certas relações probabilísticas entre palavras e categorias de assunto e destas relações predizer a categoria à qual um documento contendo a palavra pertença.

Outro trabalho em análise estatística foi o de Wallace (1965) [27]. Ele propôs levar-se em conta as palavras muito comuns num texto tais como preposições, conjunções e artigos, a fim de estruturar a análise estatística dos termos usualmente lembrados como "mais significativos"; ele demonstrou que o uso destas palavras varia suficientemente em contextos diversos, de modo a permitir a diferenciação de obras em assuntos diversos. Essa parece ser uma descoberta muito importante e poderá trazer uma contribuição considerável para precisar a categorização de textos.

Sharp (1967)[35] salientou, entretanto, que as técnicas estatísticas para análise textual são inadequadas e que isto pode ser provado de várias maneiras: uma delas é pela constatação da ausência de relatórios de trabalho nesta linha de pesquisa; outra, é pelo aparecimento na literatura do assunto, de críti-

cas veementes a estas técnicas estatísticas, onde cita como exemplo uma afirmação de Kasher (1966)[20] de que "sistemas de análise de texto baseados em princípios estatísticos não têm a capacidade de sustentar um nível de eficiência acima da média".

A idéia de Luhn de verificar termos em textos em bases puramente estatísticas trouxe uma enorme esperança para todos os estudiosos nesse campo, porque apresentava uma possibilidade de resolver o problema de classificação automática, sem maiores envolvimentos com os de semântica e sintaxe. Essa idéia, porém, provou ser inaplicável e, como conseqüência, pouquíssimas pessoas estão ainda trabalhando neste problema em comparação ao número que havia iniciado estudos neste campo. Sharp[35] citou algumas das pessoas que ainda estão otimistas a respeito das possibilidades da classificação automática, como Lesk e Salton, Minsky, e Bobrow; mas ao mesmo tempo ele também citou o pessoal extremamente pessimista, como Dreyfus, Taube, Lipetz, Levien e Maron e Sparck-Jones.

Sharp externou também as opiniões de Savage (1965)[34] e Giuliano (1965)[16]. Savage levantou o problema da inexistência de um método válido (em sua opinião) para a avaliação da classificação automática ou da indexação automática; e Giuliano escreve abertamente sobre o seu "ceticismo" a respeito de sistemas automatizados "totalmente", embora ele pense que no futuro o sucesso nessa área possa acontecer, mas seguindo caminhos completamente diferentes daqueles adotados até então.

Em suas pesquisas, Sharp[35] mais uma vez salientou que uma nova tendência estaria surgindo no sentido de se reter um pouco o desenvolvimento de novos métodos e examinar os de elaboração de resumos

e indexação tradicionais com a finalidade de se isolar as características significativas desses métodos, de modo que a máquina possa ser empregada para simular a metodologia em questão, e desta maneira atingir os mesmos resultados. E Sharp destaca então Prywes (1965)[29] como exemplo de pesquisador que tem seguido essa linha em ligação com a classificação automática.

Prywes desenvolveu uma "árvore", através da criação de nós em diversos níveis, de acordo com a freqüência de descritores nas descrições dos documentos. Assim, os descritores mais freqüentes aparecem no nível mais genérico e definem a classe geral a qual todos os documentos na coleção pertencem. Ele também elaborou um método de derivação de grupos de descritores mutuamente exclusivos, o qual é muito interessante. Aliás, um método similar foi desenvolvido posteriormente por Lefkovitz[23] conforme será tratado mais adiante neste artigo.

Em 1962, Borko[5] introduziu o conceito do uso da análise fatorial para a classificação automática. Nesse estudo, ele derivou um conjunto de categorias de classificação através da análise fatorial. Ele não pôde provar, entretanto, que tal sistema de classificação de bases empíricas fosse o melhor para determinado conjunto de documentos, tanto que, numa experiência posterior,[7] ele próprio levantou uma dúvida sobre a validade de um sistema de classificação assim gerado (ver p. 158, 3º parágrafo). Como solução, propôs a elaboração de uma outra experiência onde partiria do pressuposto de que o conjunto de classes de assunto, inicialmente derivado para literatura em computação, naquele caso, não seria ótimo. Levantou a possibilidade de que os termos de indexação usados não foram derivados apropriadamente, e propôs que

um novo sistema de classificação fosse derivado baseando-se em outros termos de índice. As conclusões a que ele chegou serão posteriormente descritas neste trabalho.

No artigo de 1963, Borko [7] tentou demonstrar a relação entre o sistema de classificação e o processo para a classificação automática de documentos. Ele propôs derivar matematicamente, em bases empíricas, um conjunto de categorias (um sistema de classificação), de modo que os documentos pudessem ser classificados automaticamente; propôs também determinar a precisão da classificação através de comparação com um critério. Seus resultados foram quase idênticos aos resultados obtidos por Maron [26] numa experiência anterior, mas Maron utilizara um outro conjunto de classes de assunto e uma fórmula computacional diferente (bayesiana). Borko utilizou em sua pesquisa 405 resumos de documentos (mesmos dados de Maron) no campo de computação. Esses documentos foram classificados manualmente de acordo com as classes derivadas anteriormente, em cinco assuntos, por dois pesquisadores. Essa classificação seria usada posteriormente como base para avaliar o desempenho da classificação a ser feita automaticamente para os mesmos documentos. Os 405 documentos haviam sido anteriormente divididos em 2 grupos: experimental e para validação. O grupo experimental compunha-se de 260 documentos. Estes foram usados para a derivação do sistema de classificação de Borko, construído com base nos 90 termos de índice selecionados manualmente por Maron. O grupo para validação tinha 145 resumos, que foram postos à parte até que todos os processos estatísticos com o grupo experimental fossem desenvolvidos, e só então foram classificados, para fins de verificação da validade desses processos.

Maron [26] idealizara um conjunto de 32 classes de assunto; Borko chegou a um total de 21 classes. Ele utilizou um programa de computador [28] que contava o número de vezes que cada um dos 90 termos do índice aparecia em cada documento. Uma matriz foi construída com essas contagens de freqüência e coeficientes de correlação foram computados para cada um dos 90 termos do índice correlacionados com cada um dos outros termos. Esta matriz foi então analisada fatorialmente.

Discussões sobre o auxílio do computador à análise fatorial foram descritas por Harman [17] e Fruchter. [15]

Por fim, para classificar automaticamente, Borko seguiu os seguintes procedimentos:

a) à(s) classe(s) contendo o termo do índice é designado um valor igual ao produto do número de ocorrências da palavra no resumo, e o fator de carga normalizado da palavra na classe. Se mais de um termo de índice aparecer na categoria (classe), os produtos são somados.

b) após cada termo de índice ser considerado, a classe tendo o mais alto valor numérico é selecionada como a mais provável classificação do assunto para o documento em questão.

Os resultados foram confrontados com os de Maron. Esses últimos demonstraram ser sempre superiores aos de Borko, principalmente com os documentos do grupo experimental, o que foi surpreendente, pois este era o conjunto de dados do qual o sistema de classificação e os fatores de carga foram derivados.

Levantaram-se questões a respeito desse acontecimento, e formularam-se duas hipóteses:

1. isso aconteceu porque categorias derivadas matematicamente não eram ótimas;

2. ou porque a equação de prognósticos baseada nos fatores de carga não agiu tão discriminatoriamente como a fórmula usada por Maron.

Para achar essas respostas, duas novas pesquisas foram elaboradas e apresentadas num trabalho posterior, que será descrito abaixo.

Em suas pesquisas adicionais (1964), Borko [8] derivou os termos de índice por freqüência de ocorrência (que eram limitados a 90) e uma vez mais derivou suas classes de assuntos, que eram também 21, com ligeiras mudanças. A partir daí, ele se propôs a responder às seguintes perguntas:

Experiência nº 1

A. Utilizando-se do *esquema original de classificação*,[7] a classificação automática de documentos terá melhor resultado se realizada através de uma *equação de predição bayesiana* do que através de *contagens de fator* (factor scores).

B. Utilizando-se o *esquema modificado de classificação*, a classificação automática de documentos terá melhor resultado se realizada através de uma *equação de predição bayesiana* do que através de *contagens de fator*.

Experiência nº 2

C. Documentos serão corretamente classificados no *esquema modificado de classificação* em número significativamente maior do que no *esquema de classificação derivada*, usando tanto o processo *bayesiano* quanto o de *contagem de fator* para a classificação automática de documentos.

Com a experiência nº 1 (A-B), concluiu-se que não havia diferença estatisticamente significativa na capacidade de classificar automaticamente documentos da *equação de predição bayesiana* e das *contagens de fator*.

Mas, na experiência nº 2 (C), os resultados mostraram que um maior número de documentos foi classificado corretamente usando-se o *esquema modificado* e não o *original* e que o aumento foi estatisticamente significativo na situação de maior importância quando prognosticando a classificação dos documentos que não haviam sido examinados previamente no grupo de validação.

Deste modo, Borko chegou a conclusões muito importantes:

1. É possível derivar matematicamente um conjunto de classes de assuntos (categorias de classificação) que sejam descritivas das dimensões mais importantes do conteúdo de uma população de documentos. Além disso, estas dimensões são relativamente estáveis, enquanto a população de documentos que as originou for igualmente estável e imutável.

2. A classificação automática de documentos é realizável e pode ser feita através da utilização tanto de processos *bayesianos* como de *contagem de fator*.

3. Se a classificação automática de documentos for adotada, obter-se-iam resultados superiores através da utilização de classes de assuntos derivadas matematicamente com base em análise estatística de palavras nos documentos e técnicas de indexação estatística.

Foi também ressaltado que as classes de assuntos foram derivadas com sucesso pela análise fatorial para relatórios de psicologia e para literatura.

Faz-se uma observação muito interessante: até agora, a classificação automática tem sido comparada, para avaliação, com a classificação humana e com esta comparação, o nível de desempenho da classificação automática tem alcançado no máximo 55% de precisão. Quem prova, entretanto, que a classificação humana é o padrão perfeito? A classificação automática deveria na realidade ser testada através de sua atuação na recuperação de documentos, porque essa é a verdadeira finalidade de qualquer sistema de classificação, isto é, levar a informação ao usuário.

Num artigo mais recente (1966), Borko [6] apresenta novamente um relato de sua classificação automática baseada em análise fatorial. Categorias (classes) são geradas por técnicas de correlação aplicadas a um número de "termos-etiqueta" ("tag terms") extraídos do corpo dos documentos (resumos, no caso). A facilidade de revisão constante das classes e redesignação dos documentos a elas, à proporção que a área de assunto sofre mudanças, é apontada como uma vantagem evidente da classificação automática.

Conforme Sharp,[35] Zavala e Van Cott (1966),[42] também trabalharam com análise fatorial. Eles tentaram classificar revistas científicas pelas áreas de assunto abordadas pelas mesmas. Simultâneamente, Doyle e Blankenship (1966) [13] apresentaram uma nova abordagem do problema da classificação automática. Dissertaram sobre a distinção entre o processo usado por seres humanos em geral na designação de documentos em esquemas de classificação já existentes e o processo usado pelos matemáticos de comparar cada documento com os outros documentos na mesma coleção a fim de gerar as classes automáticamente. Eles também abordaram o problema dos custos do computador.

Sharp [35] destaca também Williams (1966),[28] em sua tentativa de identificar os parâmetros que determinam a atuação da classificação automática, baseada em medidas de semelhança entre conjuntos de palavras nos documentos e os conjuntos que definem as classes. Ele descobriu que o fator mais significativo era o tamanho da amostragem de documento usada na definição inicial de cada classe. Uma observação importante é a que a técnica utilizada aqui era efetiva em níveis diferentes de classificação, e Williams afirma que o sucesso com este método de função discriminativa múltipla atingiu a alta marcação de 92%.

Numa revisão anterior à de Sharp, Baxendale (1966)[3] salienta a opinião de Needham [27] sobre todas as abordagens estatísticas (tais como as de Lefkovitz,[22] Dale & Dale,[10] Doyle,[11] Williams)[41] à derivação de classes — "eles são necessariamente empíricos porque não há um suporte teórico sólido para classificação automática".

A sugestão de Richmond [30] é também de grande importância; ele diz que uma condição indispensável para se chegar à verdade é que as classes, não importa como determinadas, "devem ser reconhecidas e denominadas a fim de se assegurar uma fácil comunicação da entidade da classe de pessoa para pessoa". Até então, nenhuma técnica de classificação de bases matemáticas havia atingido esse padrão.

Na mesma revisão, os trabalhos de Doyle,[11,12] e Needham [27] são discutidos. Eles tentaram estabelecer critérios estatísticos para agrupar palavras e termos de índice em classes de termos "associados". Apesar de ter havido progresso nas técnicas computacionais para implantação do conceito, a qualidade dos resultados e o desenvolvimento de uma avaliação objetiva parecem estar estáticos.

Baxendale [3] descreve também outras pesquisas realizadas com a idéia de sinonímia ou quase-sinonímia (isolamento de uma única relação associativa para estudo). Vários cientistas trabalharam com este conceito de duas palavras, embora usando amostragens de dados diversos (títulos, definições, etc.) e também peculiaridades contextuais diferentes (proximidade, etc.). Entre estes pesquisadores, temos: Edmundson, [14] Lewis,[24] Rubenstein e Goodenough,[31] e Sparck-Jones.[36]

Entre as contribuições mais significativas ao campo da organização automática da informação estão os estudos de Salton. Sobre ele, Lancaster e Gillespie (1970),[21] dizem: "seu trabalho e o de seus colegas em Cornell continua a gerar tanto quantidade como qualidade de descobertas na elaboração de sistemas de informação automáticos". O sistema SMART de Salton é reconhecido como o sistema de recuperação automática mais sofisticado que se encontra atualmente em funcionamento.

Salton (1968)[33] trabalhou dentro da área de classificação automática na análise de técnicas para a geração de conjuntos ou grupos. Estas técnicas foram as seguintes:

1. análise matricial de «eigenvalues»;
2. análise fatorial;
3. análise de classes latentes;
4 teoria dos conjuntos (clumps);
5. agrupamentos densos.

Ele avaliou alguns desses métodos em função da sua eficiência de recuperação. Através dessa avaliação, ele verificou, entre outras coisas, que métodos de análise fatorial, tais como os de Borko [7] e Bonner (1964),[4] não são muito bem sucedidos, porque, em geral,

notou-se que muitos documentos não puderam ser prontamente classificados nas respectivas classes de assunto.

Citou também uma pesquisa que comparou duas operações de agrupamento relativamente fracas (as de Bonner e Rocchio), e descreveu posteriormente outras operações desse tipo, de pouco valor, tais como a que ele próprio havia usado no sistema SMART, e outra originalmente atribuída a Needham.

Trazendo talvez uma última mensagem de otimismo, Lancaster e Gillespie,[21] numa revisão da literatura de 1969, salientaram o que parecia ser um redespertar do interesse na elaboração de sistemas englobando indexação, classificação ou pesquisa automáticas e eles atribuiram esse redespertar (ao menos parcialmente) à disponibilidade de equipamento para processamento "on-line". Lancaster e Gillespie também observaram que a única exceção nesse campo era o trabalho de Salton, cujo entusiasmo não arrefeceu em momento algum.

Citam-se ainda outros trabalhos, como os de: Armitage et al.,[1] Carrol e Soeloffs,[9] West (SPIRAL),[38] Artandi e Wolf (MEDICO),[2] Jackson,[18,19] Salesbury e Stiles,[32] e finalmente o sistema BROWSER de Williams.[40]

Uma das últimas contribuições ao campo da classificação automática foi o algoritmo desenvolvido por Lefkovitz (1969).[23] Ele elaborou um esquema de classificação em "ciência da computação" onde as descrições do documento serviram como base para a classificação (em vez de ser usada uma divisão do conhecimento feita "a priori"). Cada coleção de documentos, portanto, criou sua própria classificação baseada nas descrições de documentos de toda a coleção da biblioteca. À medida em que novos documentos

eram adicionados ao acervo, a classificação ia sendo reconstruída.

Lefkovitz construiu a sua classificação com base numa estrutura hierárquica (estrutura de árvore) na qual conjuntos de descritores (etiquetas ou palavras-chave) eram gerados cada vez que os nós da árvore e cada nó descendente (e seu conjunto de descritores associados) estavam de alguma maneira subordinados e mais específicos em relação a seu pai ou a algum nó ascendente (mais genérico) mais alto.

Sua classificação difere daquelas nas quais os descritores são designados conforme uma hierarquia semântica como a "Physics-Optics-Diffraction" do "DDC Thesaurus". Em vez disso, os descritores são colocados em conjuntos (S_1, $S_{1.1}$, etc.), não existindo pressuposições feitas sobre relações semânticas.

Nessa árvore, cada nó representa um conjunto de descritores (palavras-chave).

S = conjunto de descritores;

k = número do nó;

S_k e cada documento são representados por um conjunto de descritores associados com *um nó terminal.*

S = conjunto de descritores;

S_d d = documento.

É possível que S_d possa se encontrar num caminho (path) que tenha vários nós terminais, aos quais ele possa ser designado. Uma decisão deveria então ser tomada para a escolha do nó terminal ao qual o documento seria designado.

A árvore de classificação tem duas propriedades:

1. Cada documento é descrito por um conjunto de descritores (S_d) o qual está contido inteiramente dentro de um conjunto de nós que formam um caminho nessa árvore, do ápice ao nó terminal.

2. *Cada descritor* aparecerá *somente uma vez* dentro de um conjunto de nós do caminho. Lefkovitz associou uma unidade de máquina chamada *célula* a cada *nó terminal*. Dentro de cada célula são armazenados os dados do documento com seus descritores (chaves) que são tirados de conjuntos de nós que se localizam nos caminhos que terminam no *nó terminal.* (A célula pode ser um cilindro ou um arquivo em disco, uma série de pistas num disco de uma cabeça por pista, uma fita ou cartão de um "Armazém de Células de Dados" (Data Cell Storage), ou até um segmento de fita magnética).

Com o fim de construir esse sistema, duas árvores foram elaboradas:

1. A árvore intermediária (T_1) e
2. A árvore de classificação (T_c).

Lefkovitz observou que o seu sistema, além de possuir as propriedades de um sistema para classificação automática, poderia também ser utilizado para facilitar a pesquisa, auxiliando o usuário a ter uma visão geral, através da máquina, do material que a biblioteca possui, ou mesmo permitindo que ele formule perguntas ao sistema.

CONCLUSÃO

Como a nossa maior preocupação, nesta revisão, é a de conhecer os trabalhos em classificação automática, podemos achar interessantes as outras finalidades do sistema de Lefkovitz, mas ficaríamos mais satisfeitos se, em vez de se preocupar com estes outros usos do seu sistema, ele houvesse testado práticamente a utilização de sua classificação automática.

Parece que ele nos oferece um bom algoritmo, e é uma nova técnica, que faz renascer a esperança

dos estudiosos nesta área, uma vez que ficou provado que as demais técnicas não foram bem sucedidas quando testadas na prática, e por isto foram abandonadas.

O trabalho de Borko é também muito relevante para a área, embora não tenha ficado claro se a sua fórmula (através de análise fatorial) ou a fórmula computacional de Maron (bayesiana) é realmente a mais apropriada para a derivação automática de um sistema de classificação. (Como já foi dito anteriormente, parece que G. Salton não ficou muito satisfeito com os resultados obtidos por Borko com suas técnicas de análise fatorial).

As pesquisas de Salton continuam trazendo grandes contribuições ao campo da classificação automática e esperamos que ele ainda nos traga grandes revelações nesta área.

Gostaríamos ainda de ressaltar que só foi possível levantarmos a literatura em classificação automática até 1970 (exclusive), mas cremos que, com essa introdução, os interessados no assunto poderão continuar acompanhando os estudos realizados.

A literature review on automatic classification, presenting the different works, methods, and researchers, and pointing out among them, the ones which gave the most relevant contribution to the field of automatic classification.

BIBLIOGRAFIA

1. ARMITAGE, J. E.; LYNCH; M. F. & PETRIE, J. H. Computer generation of articulated subject indexes. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE ANNUAL MEETING, 32d, San Francisco, 1-4. Oct. 1969. *Proceedings, vol. 6: Cooperating information societies, p. 253-257.*

2. ARTANDI, SUSAN; WOLF, EDWARD, H. The effectiveness of automatically generated weights and links in mechanical indexing. *American Documentation, 20*:3 198-202, july 1969.

3. BAXENDALE, P. Content analysis, specification and control. *Annual Review of Information. Science and Technology, 1*:89-106, 1966.

4. BONNER, R.E. In some clustering techniques *IBM J. Res. Develop, 8*(1), Jan. 1964.

5. BORKO, H. The construction of an empirically based mathematically derived classification system. *Proc. Spring Joint Comput. Conf., 21*:279-289, 1962.

6. BORKO, H. Experimental studies in automated document classification. *Libr. Sci. Slant. Doc.,* (India), *3*:88-98, mar. 1966.

7. BORKO, H. & BERNICK, M. Automatic document classification. *J. ACM, 10*(2):151-62, Apr. 1963.

8. BORKO, H. & BERNICK, M. Automatic document classification. Part. II. Additional experiments. *J. ACM, 11*(2):138-51, Apr. 1964.

9. CARROLL, J.M. & Roeloffs, R. Computer selection of keywords using word-frequency analysis. *Americ. Doc. 20*(3):227-33, July 1969.

10. DALE, A.G. & DALE, N. Some clumping experiments for associative document retrieval. *Amer. Doc. 16*:5-9, Jan. 1965.

11. DOYLE, L.B. Is automatic classification a reasonable application of statistical analysis of text? *J. ACM, 12*:473-489, Oct. 1965.

12. DOYLE, L.B. *Re-expression in standardized code to improve the automatic classificability of text items. Report No TM-2213*. Santa Monica, Calif. System Development Corp., 1965. 32 p.

13. DOYLE, L.B. & BLANKENSHIP, D.A. Technical advances in automatic classification. In: BLACK, D.V.,

ed. *Proceedings of the 1966 ADI annual meeting*. Woodland Hills, Calif., Adrianne Press, 1966, p. 63-71.

14. EDMUNDSON, H.P. *Mathematical models of synonymy*. Santa Monica, Calif., Systems Development Corp., 1965. 17 p. (Preprint). Presented at the International Conference on Computational Linguistics, New York, 19-21, May 1965.

15. FRUCHTER, B. & JENNINGS, E. Factor analysis Nº 1. In: BORKO, H., ed. *Computer applications in the behavioral sciences*. Englewood Cliffs, N.J., Prentice-Hall, 1962.

16. GIULIANO, V.E. Postscript: a personal reaction to reading the conference manuscripts. In: STEVENS, M.E.; GIULIANO V.E.; HEILPRIN, L., eds. *Statistical association methods for mechanized documentation;* symposium proceedings, Washington, 1964. (Washington, D.C., U.S. Dept. of Commerce. National Bureaux of Standards Miscellanious Publication 269).

17. HARMAN, N.H. *Modern factor analysis*. Chicago, University of Chicago, 1947.

18. JACKSON, DAVID, M. Basis for an improvability measure for retrieval performance. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE ANNUAL MEETING. 32 d, San Francisco, 1-4, October, 1969. *Proceedings,* vol. 6: *Cooperating information societies,* p. 487-494.

19. JACKSON, DAVID, M. *The construction of retrieval environments and pseudo-classifications based on external relevance*. Columbus, Ohio State University, Computer and Information Science Research Center, 1969, 74 p. (Technical Report 69-3).

20. KASHER, *A*. *Data-retrieval by Computer;* a critical survey. Jerusalém, Hebrew Univ., 1966, 72 p. (Technical report nº 22 to Office of Naval Research, Information Systems Branch), (AD-631-748).

21. LANCASTER, F.W. & GILLESPIE, C.J. Design and evaluation of information systems. *Annual Review of Information Science and Technology,* 5:33-70, 1970.

22. LEFKOVITZ, D. The application of the digital computer to the problem of a document classification system. In: CHEYDLEUR, B.F., ed. *Proceedings of the Colloquium on Technical Preconditions for Retrieval Center Operations*, Philadelphia, Pt., 24-25 April, 1964. Washington, D.C., Spartan Books, 1965. p. 133-146.

23. LEFKOVITZ, D. *File structures for on-line systems*. New York, Spartan Books, 1969, p. 186-201.

24. LEWIS, P.A.W.; BAXENDALE, P.; & BENNETT, J.L. *Statistical discrimination of the synonymy antonymy relationship between words*. San José, Calif., IBM Co. Research Lab, 1965, 33 p.

25. LUHN, H.P. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev., 1*:309-17, 1957.

26. MARON, M.E. Automatic indexing; an experimental inquiry. *J. ACM, 8*:407-17, 1961.

27. NEEDHAM, R.M. Applications of the theory of clumps. *Mech. Trans.*, *8*(3,4) June- Oct., 1965.

28. OLNEY, J.C. *FEAT, an inventry program for information retrieval*. *FN*-4018. Santa Mônica, Calif., System Development Corp., 1960.

29. PRYWES, N.S. Browsing in an automated library through remote access. In: SASS, M. & WILKINSON, W., ed. *Computer augmentation of human reasoning*. Washington, D.C., Spartan Books, 1965, p. 105-130.

30. RICHMOND, P.A. Transformation and organization of information content: classification research. In: FID Congress, Washington, D.C., 10-15, October, 1965. *Abstracts*. Washington, D.C., Secretariat, FID Congress, 1965, p. 25.

31. RUBENSTEIN, H. & GOODENOUGH, J.B. Contextual correlates of synonymy. *Comm. ACM, 8*:627-633, Oct. 1965.

32. SALISBURY, B.A., JR. & STILES, H.E. The use of the B-coefficient in information retrieval. In: AME-

RICAN SOCIETY FOR INFORMATION SCIENCE ANNUAL MEETING. 32d, San Francisco, 1-4, October, 1969. *Proceedings, vol.* 6:*Cooperating information societies,* p. 256-268.

33. SALTON, G. *Automatic information organization and retrieval*. New York, Mc Graw-Hill, 1968. p. 133-50.

34. SAVAGE, T.R. The unevaluation of automatic indexing and classification/abstrac only/ In: STEVES, M.E.; GIULIANO, V.E.; HEILPRIN, L., ed. *Statistical association methods for mechanized documentation;* symposium proceedings, Washington, 1964. Washington, D.C., U.S. Dept of Commerce. National Bureau of Standards, 1965, p. 211. (National Bureau of Standards Miscellaneous Publication 269).

35. SHARP, J.R. Content analysis, specifications, and control. In CUADRA, D.A., ec. *Annual review of information science and technology, 2*:107-122, 1967.

36. SPARCK JONES, K. Experiments in semantic classification. *Mech. Trans., 8*:97-112, June-Oct., 1965.

37. WALLACE, E.M. Bank order patterns of common words as discriminators of subject content im scientific and technical prose. In: STEVENS, M.E.; Giuliano, V.E.; & HEILPRIN, L., ed. *Statistical association methods for mechanized documentation;* symposium proceedings, Washington, 1964. Washington, D.C., U.S. Dept. of Commerce. National Bureau of Standards, 1965, p. 225-29. (National Bureau of Standards Miscellanious Publication 269).

38. WEST, LESLIE E. SPIRAL. Sandia's program for information retrieval and listing. In: AMERICAN SOCIETY FOR INFORMATION SCIENCE ANNUAL MEETING. 32d, San Francisco, 1-4, October, 1969. *Proceedings, vol.* 6: *Cooperating information societies,* p. 139-49.

39. WEST, LESLIE E. SPIRAL. *Sandia's program for information retrieval and listing*. Sandia Laboratories, Albuquerque, New México, December, 1968, 86 p. (SC-RR-68-819C).

10. WILLIAMS, JOHN H., JR. & BROWSER. *An automatic-indexing on-line text retrieval system. Annual progres report*. IBM Federal Systems Division, Gaithersburg, Md., 1969, 30 p.

41. WILLIAMS, J.H., JR., *Results of classifying documents with multyple discriminant functions*. Rockville, Md., IBM, Federal Systems Div., 1965, 31 p. (AD-612-272).

42. ZAVALA, A., & VAN COTT, H.P. *A feasibility study of the factor analysis of scientific literature*. Final report. Silver Springs, Md., American Institutes for Research, 1966, 38 p. (AIR-E-92-7/66-FR) NSF Grant Nº GN-496.