

INDEX

- Additive conjoint structure, 120, 121
Additive independence, 121
Annual Review of Information Science and Technology, v, 10, 12
Association measures, *see* Measures of association
Attributes, 34
Attribute-value pairs, 56
Automatic abstracting, 14
Automatic classification, 2, 8, 29-55, 133-134
Automatic content analysis, 2, 6, 7, 136-137
Automatic document classification, *see* Document clustering
Automatic indexing, 21, 26
Automatic keyword clustering, *see* Keyword classification
Automatic text analysis, 3, 12-28
Automatic thesaurus, *see* Keyword classification
Averaging techniques, 101-103
- Balanced tree, 72
Binary tree, 72
Boolean search, 59, 60, 81-83
Bottom-up search, 89
Bump table, 76
- Cellular multi-lists, 63-65
Centroid, *see* Cluster representative
Citation graph clustering, 52
Classification (definition), 29
Classification algorithms, 39, 42-50
Classification methods, 34-36, 38-45
Classification programs, 51
Clique, *see* Maximal complete subgraph
Clumps, 36, 42
Cluster hypothesis, 37-38
Cluster methods, *see* Classification methods
Cluster profile, *see* Cluster representative
Cluster representative, 42, 43, 44, 48, 85-87
Cluster time dependence, 49-50
Cluster-based retrieval, 38, 47, 87-89
Clustered file, 8, 78
Clustering, *see* Classification methods, Classification algorithms
Coarse clustering, 49
Collision handling methods, 75-86
Combination function, 124
Common words, *see* Stop words
Composite measures, 105
Computer-aided instruction, 138
Conflation, 15-20
Conflation algorithm, 20
Connected component, 40, 46
Connectedness, 118
Connection matrix, 46, 47
Content sensitive suffix removal, 15
Contingency table, 99
Control parameter, 101
Controlled vocabulary, 21, 24
Co-occurrence, 7, 24, 25
Co-ordination level, 83
Core custering, 49
Correlation measure, *see* Measures of association
Cosine coefficient, 32, 34
Cosine correlation, 84

INDEX

- Cranfield II, 10
Cumulative frequency distribution, 126-127
- Data retrieval systems, 1, 138
Data simplification, 35
Data structure, *see* File structures
Decision rule, 88
Decomposable structure, 121
Decreasing marginal effectiveness, 122
Degenerate trees, 72-73
Dendrogram, 45, 65, 71
Diagnosis (definition), 29
Dice's coefficient, 32
Directory, 58
Dissimilarity coefficient, 33, 45, 46
Dissimilarity matrix, 46
Document clustering, 8, 30, 36, 38, 47-50
Document frequency weighting, 23
Document representative, 5, 13, 15, 20, 25
Doubly chained tree, 68-70
Dynamic classification, 48-49
- E*-measure, 324
Effectiveness, 2, 9, 96, *see also* Measurement of effectiveness
Efficiency, 2, 9
Empirical distribution function, 126
Empirical ordering, 117
Essential component, 120
Evaluation, 1, 3, 8, 95-132, 135
Exhaustivity, *see* Indexing exhaustivity
Expected search length, 109-113
Experimental information retrieval, 1, 2, 48
- Fallout, 100, 106
Feedback, 6, 90-93
Field, 57
File organisation, *see* File structures
File structures, 3, 56-80, 134
Fixed increment error correction, 91
Formal semantics, 13
Frequency of occurrence, 7, 13, 14, 23, 24
Fuzzy sets, 52
- Generality, 100
Graph theoretic cluster methods, 39-41
- Hash addressing, 74-87
Hashing function, 75-76
Heuristic cluster methods, 42-44
Hierarchic classification, 44-48
Hierarchic cluster methods, 41-48
High-frequency words, *see* Stop words
- Independence, 119
Index language, 20-21
Index language specificity, 22
Index term, 20
Index term weighting, 22-24
Indexing, 4, 20-22
Indexing exhaustivity, 22-23
Indexing specificity, 22-23
Index-sequential file, 60-63
Information retrieval (definition), 1
Information retrieval system, 1, 5
Information structure, 6, 7
Interaction term, 121
Interactive search formulation, 89-90
Interpolation, 103-104
Intersubstitutibility, *see* Keyword substitution
Inverse document frequency weighting, 23
Inverted file, 8, 59-60
- Jaccard's coefficient, 32
- K*-list, 58
K-pointer, 58
Key, 58
Keyword, 7, 20
Keyword, classes, 25, 41
Keyword classification, 24-25, 38, 41
Keyword co-occurrence, *see* Co-occurrence
Keyword frequency, *see* Frequency of occurrence
Keyword substitution, 25
- Linguistics, 12, 13
Log precision, 116
Logical relevance, 98
Low-frequency words, *see* Rare words
- Macro-evaluation, 101-104
Matching function, 42, 43, 83-85
Maximal complete subgraph, 40
Maximally linked document, 86
Measurement of effectiveness, 2, 8, 99-130
Measures of association, 7, 31-34

- Measures of effectiveness, 8, 9, 99-101
MEDLARS, 21, 24, 90
MEDLINE, 6, 90
MEDUSA, 6, 90
Micro-evaluation, 101-102
Mini-computer, 138
Minimal premiss set, 98
Monothetic, 35
Multi-level classification, 43
Multi-lists, 63
- Natural classification, 30
Nearest neighbour classification, *see* Single-link
Negative dictionary, *see* Stop list
Networks, 138
Non-parametric tests, 128-129
Normalisation of text, 26
Normalised association measures, 32
Normalised precision, 114-115
Normalised recall, 113-114, 115
Normalised symmetric difference, 33, 34, 116-117, 124
Numerical relational structure, 121
Numerical representation, 121-122
- Operating characteristic, 107-108
Operational information retrieval, 1, 6, 48
Optimal query, 92
Order dependence, 44
Order independence, 39
Ordered classification, 36
Overlap coefficient, 32
- Pointer, 57
Polythetic, 35
Post-coordination, 20
Precision, 8, 9, 22, 97, 99-101, 106
Precision-recall curve, 100
Pre-coordination, 20
Presentation of experimental results, 125-127
Probabilistic indexing, 26
- Query representative, *see* Document representative
Question-answering systems, 1, 98, 137
- Rag-bag cluster, 43, 44
Rank recall, 116
Rare words, 15
Recall, 8, 9, 22, 97, 99-101, 106
- Reclassification, 49
Record, 56
Relational structure, 118
Relative importance of recall and precision, 123
Relevance, 4, 5, 8, 97-99
Relevance feedback, 90-93
Representation theorem, 120-122
Restricted solvability, 120
Retrieval effectiveness, *see* Measures of effectiveness
Ring structure, 65-67
Rocchio algorithm, 43
- Scale, 121
Scatter storage, 74-77
Search strategies, 3, 81-94, 135
Sequential file, 7, 59
Serial search, 38, 84
Sign test, 128-129
Significance tests, 8, 127-129
Similarity coefficients, 32
Similarity matrix, 40
Similarity matrix generation, 50
Simple matching coefficient, 32, 83
Simple ordering, 110
Simulation, 135
Single-link, 41, 45-47
Single-link algorithm, 46, 47
Single-pass algorithm, 43
SMART, v, 21, 30, 51
SNOB, 50
Specificity, *see* Index language specificity, Indexing specificity
Specificity weighting, 23
Stability of classification, 39
Standard recall values, 101
Statistical significance, *see* Significance tests
Stemming, 15-20
Stop list, 15, 16-17
Stop words, 14, 15-17
Stopping rule, 88
Stratified hierarchic cluster methods, 47-48
Suffix stripping, 15-20
Suffixes, 15, 18-19
Swets model, 105-109
- Term, *see* Keyword
Term classes, *see* Keyword classes
Term clustering, *see* Keyword classification
Term co-occurrence, *see* Co-occurrence

INDEX

- Term frequency, *see* Frequency of occurrence
- Term substitution, *see* Keyword substitution
- Terminal node, 71
- Theoretical soundness of classification, 39
- Theory of measurement, 117
- Thesaurus, 24
- Thomsen condition, 120, 125
- Threaded list, 67-70
- Top-down search, 89
- Trainable pattern classifier, 91
- Transitivity, 118
- Trees, 70-74
- Triangle inequality, 33
- Typical document, 86
- Uncontrolled vocabulary, 21
- Uniqueness of single-link, 48
- Updating classifications, 49-50
- Weak ordering, 110, 118
- Weighting schemes, 23
- Wilcoxon matched pairs test, 128
- Word frequency, *see* Frequency of occurrence
- Zipfian distribution, 13-14, 23, 26
- Zipf's law, 13-14