UNIVERSIDADE DE SÃO PAULO

A DETAILED SCHEMATIC STRUCTURE OF RESEARCH PAPER INTRODUCTIONS: AN APPLICATION IN SUPPORT-WRITING TOOLS

SANDRA M. ALUÍSIO
OSVALDO N. OLIVEIRA JR.

No. 26

NOTAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2577

A DETAILED SCHEMATIC STRUCTURE OF RESEARCH PAPER INTRODUCTIONS: AN APPLICATION IN SUPPORT-WRITING TOOLS

SANDRA M. ALUÍSIO
OSVALDO N. OLIVEIRA JR.

No. 26

NOTAS DO ICMSC
Série Computação

São Carlos
Mai./1996

A Detailed Schematic Structure of Research Paper Introductions: An Application in Support-Writing Tools

Sandra M. Aluisio
University of Sao Paulo
Department of Computer Science
CP 668, 13560-970
Sao Carlos, SP, BRAZIL
e-mail: sandra@icmsc.sc.usp.br,
sandra@meru.uwyo[1]

Osvaldo N. Oliveira Jr.
University of Sao Paulo
Institute of Physics of Sao Carlos
CP 369, 13560-970
Sao Carlos, SP, BRAZIL
e-mail: chu@ifqsc.sc.usp.br

Abstract

Corpus analysis of 54 naturally occurring introductions of Physics papers led to a detailed schematic structure comprising eight components that are realized linguistically through 30 rhetorical strategies. A strategy is made up of two or three rhetorical messages out of a set of 45 messages. Such a detailed analysis was required for employing the case-based reasoning approach in developing writing tools aimed at assisting non-native English users. The introductions and related rhetorical structures formed the case base, with cases being easily retrieved and adapted through revision rules.

1. Introduction

The reuse of linguistic material, acquired manually or using semi-automatic tools in a corpus, has been employed in various types of systems. These include report generators [Kukich-83, Smadja-91, Buchanan-92]; case-based letter generators [Pautler-94]; and hypertext-based support systems for software documentation [Born-92]. For our purposes, linguistic material is to be reused for assisting non-native English writers in preparing first drafts of scientific paper Introductions. We decided to address Introductory sections first, not only because it is one of the most difficult parts to write but also because of its interesting role as public relations [Gosden-93] and its strategic place in the paper [Swales-90].

A number of works have been published on the schematic structure of Introductory Sections of experimental research papers written in English. Some of these present models of the structure [Swales-90, Weissberg-90]; others verify the validity of the models for various research areas [Wood-82, Crookes-86]; or for different cultures [Taylor-91]. However, in order to apply case-based reasoning (CBR) (see e.g. Mantaras-95) in the design of writing tools we had to perform a more refined corpus analysis. This was needed because although the global structure of a text may possess a relatively well-defined schema, as suggested by Swales and Weissberg, its more detailed structure can be organized in different ways. Moreover, text reuse requires an indexing vocabulary sufficient to discriminate among texts

[1] Currently, the author is visiting the University of Wyoming, Department of Computer Science.

on the basis of the characteristics relevant to the purpose for which the text is being retrieved. The corpus analysis provided this precise vocabulary to apply the CBR approach for the task of drafting technical papers by non-native speakers.

This paper deals primarily with the corpus analysis on Introductions of papers on experimental research. Emphasis is given on the acquisition phase used to launch out the case-based approach for supporting the writing up. The methodology and the results of the analysis are presented in Section 2. Section 3 illustrates briefly the representation of rhetorical structures in the form of cases employed in the writing tool. This tool utilizes the case-based approach for modelling the various stages of the writing process, i.e. planning, composing and revising [Hayes-80]. The purpose is to improve the cohesion and coherence of the introduction draft. The paper concludes by identifying limitations and areas for further research.

2. Corpus Analysis of Introductory Sections

2.1 Selection of the Corpus

Introductions were taken from two journals dedicated to Physics and Materials Science. Thirty-three (33) papers were ramdonly selected from the 1992-1994 issues of Physical Review Letters (PRL) with the only restriction that the papers be produced by English native speakers. PRL was chosen not only because it is one of the most prestigious journals but also because papers must fulfill very stringent requirements, in particular with regards to being of general interest for the Physics community. Twenty-one (21) papers were taken from a special issue of the 1992 edition of Thin Solid Films (TSF), dedicated to organized thin films. The choice of papers in a more specific area helped in analysing distinct strategies employed in writing an Introduction. Again, all papers selected from TSF were written by English native speakers. The selected papers from both TSF and PRL are short (3 or 4 pages) and are highly standardized.

2.2 Objectives and Methodology

We had three objectives in mind in carrying out the corpus analysis: i) identify the different types of information (rhetorical messages); ii) how these messages are realized linguistically; iii) what types of cohesive devices are used for grouping messages into a cohesive passage. The analysis was carried out manually in each of the 54 Introductions, consisting of three steps:

1. Pre-processing: the Introduction text was subdivided in sentences which were numbered.

2. Identification of the rhetorical messages within the clauses and sentences. The non-factual linguistic material was highlighted to be reutilized.

3. Identification of patterns of rhetorical organization for linking the rhetorical messages. The patterns for building paragraphs were identified by recognizing mainly the syntactic features such as adjectives and adverbs for signaling temporal relations, logical connecters, structural parallelism and anaphoric and cataphoric references among sentences. The patterns led to the rhetorical strategies that were named according to the component in which they appeared.

2.3 Results from the Analysis

A Detailed Schematic Structure for an Introduction

We observed eight components which overall corroborate the Swales and Weissberg models: setting (S), review of the literature (RL), gap (G), purpose (P), methodology (M), main results (R), value of the research (V) and layout of the paper (L). Our model is more precise than those of Swales and Weissberg since the components are subcategorized in formalized rhetorical strategies. Figure 1 illustrates these results in the form of a schematic structure for Introductions in Experimental Physics. The order of appearance of the various components is most likely to be that shown in Figure 1, but other possibilities exist.

C1: Setting
S1 Introducing the research topic from the research area
S2 Familiarizing terms, objects, or processes
S3 Arguing about the topic prominence
C2: Review
S1 Historical review
S2 Current trends
S3 General to particular ordering for citations
S4 Progress in the area
S5 Requirements for the progress in the area
S6 State of the art
S7 Compounding review of the literature and their gaps
S8 Citations grouped by approaches
C3: Types of Gap
S1 Unresolved conflict or problem among previous studies
S2 Restrictions in previous works
S3 Raising questions
C4: Purpose
S1 Indicating the main purpose
S1A Solving a conflict among authors
S1B Presenting a novel approach, or methodology, or technique
S1C Presenting an improvement in a research topic
S1D Presenting an extension of a previous author's work
S1E Proposing an alternative approach
S1F Presenting a comparative work
S2 Specifying the purpose

S3 Introducing more purposes

C5: Methodology

S1 Listing criteria or conditions

S2 Describing materials and methods

S3 Justifying choices for methods and materials

C6: Main Results

S1 Presenting/Indicating results

S2 Commenting on the results

C7: Value of the Research

S1 Stating the value of the work

C8: Layout of the article

S1A Outlining the parts of the article

S1B Listing issues to be addressed

Figure 1: Main Components, C, and Strategies, S, comprising the detailed schematic structure of an Introduction. The strategies are numbered in each component; those with letters following the number indicate mutual exclusion.

The introductions analysed were short and therefore the components value (V), methodology (M) and layout (L) were usually absent. There were 37 different combinations of the components out of the 54 cases. Rather than demonstrating lack of conventionalization, this result shows that the number of introductions analysed is too small for repetition of such a detailed structure to occur. Indeed, when the case-based approach was employed using the corpus it was clear that the case base should be enlarged.

Approximately 80% of the introductions have the following form: [S RL G P R], with 24 out of the 54 introductions being sublists of this form. There are other 19 cases which either include V, M or S in the order indicated in Figure 1 or that differs from the canonical form by the repetition of components (for instance, appearance of a specific RL after G). The remaining 11 introductions bring specificities of their own. Important among these are the introductions starting with the purpose (P), in which the authors put great emphasis on a particular goal of their research that should be readily recognized by readers as being very important. Only one introduction presented a very complex structure [RL P R M R RL G R S] which is characteristic of longer introductions.

There are several possible rhetorical strategies for building paragraphs or text passages in scientific writing. For introductions, the most employed ones include the chronological organization of topics, analysis of cause-effect, contrasts, lists, topic organization from general to specific, illustrations, etc. [Trimble-85, Turk-89, Huckin-91]. Thirty (30) rhetorical strategies[2] were identified; each of them was linguistically realized employing two or three rhetorical messages. Some of the strategies possess common characteristics and could be generalized. But they receive distinct names for helping the user in the gathering of features as they appear in different components.

[2] They were based on works by [Bramki-84, McKeown-85, Trimble-85, Maybury-91, Flowerdew-92, Shaw-92].

Forty-five (45) types of rhetorical message were identified. The linguistic material of each sentence that could be reused (non-factual information) for the writing of other texts was highlighted. Some sentences were obviously discarded as they conveyed too specific a statement and/or generally lacked rhetorical markers (or cohesive devices) which made them too dependent on the context.

Examples of Rhetorical Strategies and Related Messages

The following strategy[3] from the Setting component is one of the most used strategies in our corpus for starting the Introduction.

Arguing about the topic prominence:
Claiming about topic prominence
Familiarization*
Support+

An example can be seen in Figure 2 where "Claiming about topic prominence" is instantiated with "A great deal of interest has recently been stimulated by the use of organic materials in electroluminescent (EL) devices [1]." The strategy "Claiming about topic prominence" is defined as:

(*Claim relevance/ Claim currently active/ Claim well-established*).

The introduction presented in Figure 2 utilizes the "Claim currently active" message for the "Claiming about topic prominence strategy" as it makes use of a time adverb — "recently" — indicating the research is currently active.

The Support strategy is definided as:

(Motivation/ Cause/ Result/ Purpose/ Evidence/ Particularization/ Exemplification).

The "Familiarization" strategy can appear as a main strategy as well and its definition is given below.

Familiarizing terms, objects or processes:
Familiarization+

where "Familiarization" is defined as: (**Definition/ Description/ Classification**).

[3] The notation [x] indicates that strategy (or message) "x" is optional, x* indicates occurrence of "x" zero or more times, x+ indicates one or more occurrences of "x" and (x/y) indicates appearance of "x" or "y". Strategies appearing in bold may be refined in other strategies or rhetorical messages. Rhetorical messages appear in italic.

"Definition" is defined as: (*Formal Definition/ Semi-Formal Definition/ Substitution/ Stipulation/ Definition by Exemplification/ Definition by Particularization*);

"Description" is considered as: (*Physical description/ Function description/ Process description);*

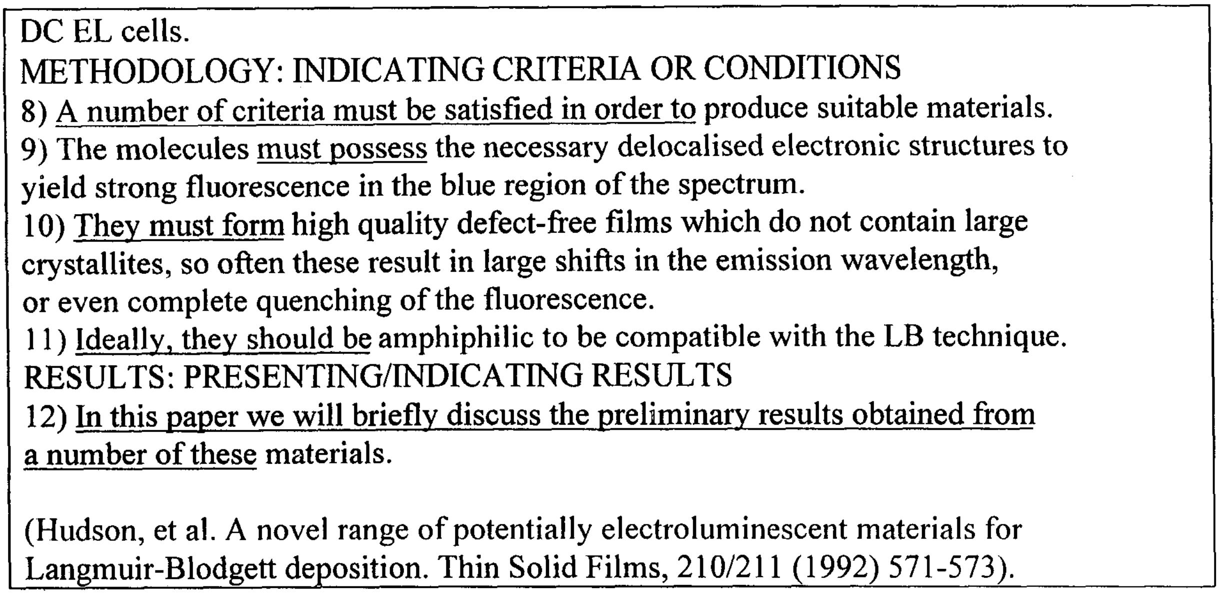
and "Classification" as: *(Complete classification/ Partial classification).*

Some types of definition (by substitution and semi-formal) and descriptions are encountered in various parts of an Introduction. But the familiarization strategy appears mainly in the Setting component when terms must be defined, the extent of which depends on the target readers.

Listing Criteria or conditions:
[Head of the List]
(Criteria or Conditions)+

This strategy follows the enumerative list pattern which may (or may not) include an initial message indicating the intention of listing ("Head of the List") which is followed by a list of criteria or conditions. Figure 2 illustrates an Introduction making use of this startegy.

SETTING:ARGUING ABOUT THE TOPIC PROMINENCE
1) A great deal of interest has recently been stimulated by the use of organic materials in electroluminescent (EL) devices [1].
2) Organic molecules can be engineered to possess specific functional properties, offering the possibility of obtaining intense fluorescence which can be tuned to a particular wavelength.
GAP:RESTRICTIONS IN PREVIOUS APPROACHES
3) However, the fabrication of EL devices with bright blue emission has proved difficult owing to the bathochromic shifts in emission wavelength which often occur between solution and film spectra.
REVIEW:PROGRESS IN THE AREA
4) Much previous work has been directed towards vacuum deposited films and significant progress has been made by incorporating charge transport layers into multilayer EL cells [2, 3].
5) Since it provides precise control over film thickness and a high level of molecular ordering within each layer, the Langmuir-Blodgett (LB) technique is well suited to this application.
6) Blue electroluminescence in a Langmuir film was first reported in 1980 using anthracene as an emitter layer [4].
PURPOSE: PRESENTING A NOVEL APPROACH, METHODOLOGY, OR TECHNIQUE
7) In this research we will examine the use of novel materials to be used in blue



DC EL cells.
METHODOLOGY: INDICATING CRITERIA OR CONDITIONS
8) A number of criteria must be satisfied in order to produce suitable materials.
9) The molecules must possess the necessary delocalised electronic structures to yield strong fluorescence in the blue region of the spectrum.
10) They must form high quality defect-free films which do not contain large crystallites, so often these result in large shifts in the emission wavelength, or even complete quenching of the fluorescence.
11) Ideally, they should be amphiphilic to be compatible with the LB technique.
RESULTS: PRESENTING/INDICATING RESULTS
12) In this paper we will briefly discuss the preliminary results obtained from a number of these materials.

(Hudson, et al. A novel range of potentially electroluminescent materials for Langmuir-Blodgett deposition. Thin Solid Films, 210/211 (1992) 571-573).

Figure 2: One of the introductions of the Case Base. Some fragments are underlined which correspond to the reusable parts of the text.

3. The Application

The detailed schematic structure, the notation used in defining the rhetorical strategies, and the 54 rhetorical structures of the introduction corpus were used in a case-based system for assisting non-native English speakers to write scientific texts. These knowledge sources were framed up in the stages of the writing process: planning, composing and revising. Accordingly, the user follows a three-step procedure: i) gathering of features, in which the user selects from several menus the features intended for his/her introduction; ii) selection of the best-match case, following the case recovery by the system; iii) revision on the selected case.

For recovering cases, three ways of pattern matching between requisition and cases are used: perfect match (equal lists), proper undermatch (sublist) and non-proper undermatch (intersection). The tool selects cases to be returned to the user by employing these three metrics, which are basically related to the degree of certainty on the part of the user about the order of the components and strategies: "sure about the order", "some doubts", "many doubts". The base contains the 54 introductions analysed, each introduction plus its rhetorical structure consisting in a case (see Figure 2 illustrating one of the cases and its corresponding representation in Prolog in Figure 3). For revision, four operations were envisaged: i) changes in the lexical and syntactic material of the messages; ii) changes applied to the selected strategies recovering similar ones; iii) addition of messages to a specific strategy; iv) deletion of messages, the opposite operation to iii). The operations ii), iii) and iv) derived directly from the notation for rhetorical strategies in Section 2: a message may be an alternative to another message, may occur one or more times, zero or more times, and may be optional. The use of similarity metrics employed for recovering cases (whole Introductions) can also be used for recovering strategies and ultimately messages from the phrasal lexicon, even if they belong to different cases. Operation i) was designed for lending paraphrasing power to the tool and is realized by applying the perfect match rule to the entries in the phrasal lexicon. The details of computational implementation are published in [Aluisio-95].

```
case(tsf7,
[c(setting,s(argumenting_about_the_topic_prominence,
    [m(claim_currently_active2,tsf7,1),m(motivation,tsf7,2)])),
c(gap,s(unresolved_conflict_or_problem_among_previous_studies,
    [m(conflict_or_problem,tsf7,3)])),
c(review,s(progress_in_the_area,
   [m(claim_progress,tsf7,4),m(cause,tsf7,5),m(claim_relevance,tsf7,5),
    m(particular_fact,tsf7,6)])),
c(purpose,s(presenting_a_novel_approach_or_methodology_or_technique,
   [m(novel_approach, tsf7,7)])),
c(methodology,s(listing_criteria_or_conditions,
   [m(head_of_the_list,tsf7,8), m(criteria_or_conditions,tsf7,9)])),
c(results,s(presenting_results,
   [m(indicating_results,tsf7,12)]))],_).
```

Figure 3: Rhetorical features represented in Prolog. A case is represented by the Prolog structure case(Case_Name, ListofComponents, ListofPragmaticFeatures). Each of the components in the ListofComponents is represented by c(Component_Name, Strategy). The strategies have the Prolog structure s(Strategy_Name, ListofMessages) which should include a few messages.The last component of the case structure is not being used in developing the present tool.

4. Further Work

In order to reuse non-factual linguistic material in a software writing tool the detailed schematic structure must be determined for the type of text under consideration. For papers on Experimental Physics, introductions were shown to consist of eight components that are realized linguistically through 30 types of rhetorical strategies. Each strategy, in its turn, is made up of 2 or 3 out of a set of 45 rhetorical messages. The corpus analysis provided linguistic material for the two knowledge sources: the case base and the revision rules by which a case can be adapted for the user own needs. The prototype demonstrates the feasibility of the case-based reasoning approach for developing software tools aimed at assisting non-native English users, with cases being easily retrieved and interactively adapted.

One of the main limitations of the tool is that the case base must be considerably extended before meaningful tests on precision and recall can be performed. Nevertheless, preliminary results with target users have shown that the tool allows users with a limited command of technical English to generate a cohesive and coherent first draft of introductory texts, although the drafts still have intra-sentential problems related to lexical and grammatical mistakes on the inserted linguistic material. This seed work may also be extended in several ways for users of other areas of research and for other sections of a paper. In order to alleviate the burden of knowledge acquisition we are now employing a semi-automatic tool based on pattern matching for enlarging the case base and the phrasal lexicon.

References

[Aluisio-95] Aluisio, S.M. and Oliveira Jr. O.N. A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users. Lectures Notes in Artificial Intelligence 1010. pp. 121-132, 1995.

[Born-92] Born,G. A Hypertext-Based Support Aid for Writing Software Documentation. In Computers and Writing - State of the Art, P. O'Brian-Holt and N.Williams (eds), Kluwer Academic Publishers, Dordrecht, pp. 266-277, 1992.

[Bramki-84] Bramki,D. and WilliamsR..Lexical Familiarization in Economics Text, and its Pedagogic Implications in Reading Comprehension. Reading in a Foreign Language, Vol 2, Nro 1, Spring, pp.169-181,1984.

[Buchanan-92] Buchanan,R.A. Textbase Technology: Writing with Reusable Text. In Computers and Writing - State of the Art, P. O'Brian Holt and N.Williams (eds), Kluwer Academic Publishers, Dordrecht, pp. 254-265, 1992.

[Crookes-86] Crookes, G. Towards a Validated Analusis of Scientific Text Structure. Applied Linguistcs, Vol 7 (1), pp. 57-70, 1986.

[Flowerdew-92] Flowerdew,J. Definitions in Science Lectures, Applied Linguistics, Vol 13 (2), June, pp. 202-221, 1992.

[Gosden-93] Gosden, H. Discourse Functions of Subject in Scientific Research Articles. Applied Linguistics, Vol 14 (1), pp. 56-75, 1993.

[Hayes-80] Hayes, J.R. & Flower, L.S. Identifying the Organization of Writing Processes. In L.W. Gregg & E. R. Steinberg (eds.), Cognitive Processes in Writing, Hillsdale, N.J. Erlbaum. 1980, pp. 3-30.

[Huckin-91] Huckin,T.N. and Olsen,L.A. Technical Writing and Professional Communication for Nonnative Speakers of English. McGraw-Hill, Inc, 1991.

[Kukich-83] Kukich,K. Knowledge-Based Report Generation: A Knowledge Engineering Approach to Natural Language Report Generation. PhD Thesis, University of Pittsburg, 1983.

[Mantaras-95] Mantaras,R.L. and Plaza,E. Case-Based Reasoning. In The Newsletter of the European Network of Excellence in ML, pp. 29-37, Special Issue, September, 1995.

[Maybury-91] Maybury,M.T. Planning Multisentential English Text Using Communicative Acts PhD Thesis -- Tech. R. 239, University of Cambridge, 1991.

[McKeown-85] McKeown,K.Discourse Strategies for Generating Natural-Language Text Artificial Inteligence 27, pp. 1-41, 1985.

[Pautler-94] Pautler,D. Planning and Learning in Domains Providing Little Feedback. In AAAI Fall Symposium on Planning and Learning Notes'94, 1994.

[Shaw-92] Shaw,P. Reasons for the Correlation of Voice, Tense, and Sentence Function in Reporting Verbs, Applied Linguistics, Vol 13(3), September, pp. 302-319, 1992.

[Smadja-91] Smadja,F.Retrieving Collocational Knowledge from Textual Corpora. An application: Language Generation. PhD Thesis, Computer Science Department, Columbia University, 1991.

[Swales-90] Swales,J. Genre Analysis - English in academic and research settings. Cambridge University Press, 1990.

[Taylor-91] Taylor, G. and Tingguang, C. Linguistic, Cultural and Subcultural Issues in Contrastive Discourse Analysis: Anglo-amarican and Chinese Scientific Texts. Applied Linguistics, Vol. 12 (3), September, pp. 319-336, 1991.

[Trimble-85] Trimble,L. English for science and technology: a discourse approach. Cambridge University Press, 1985.

[Turk-89] Turk,C. and Kirkman,J. Effective Writing. E. and F.N. SPON, London, 1989.

[Weissberg-90] Weissberg,R. and Buker,S. Writing up Research - Experimental Research Report Writing for Students of English. Prentice Hall Regents, 1990.

[Wood-82] Wood, A.S. An Examination of the Rhetorical Structures of Authentic Chemistry Texts.Applied Linguistics, Vol. 3 (2), pp. 121-143, 1982.

NOTAS DO ICMSC

SÉRIE COMPUTAÇÃO

025/96 NUNES, M.G.V.; HASEGAWA, R.; KAWAMOTO, S.; OLIVEIRA, M.C.F. DE; TURINE, M.A.S.; GHIRALDELO, C.M.; OLIVEIRA JR., O.N.; RIOLFI, C.R.; SIKANSKI, N.S.; MARTINS, T.B. - Style and grammar checkers for brazilian portuguese.

024/96 FORTES, R.P.M. - Uma ferramenta orientada a links para avaliação de hiperdocumentos.

023/96 CANSIAN, A.M.; MOREIRA, E.S.; MAURO, R.B.; MORISHITA, F.T.; CARVALHO, A.C.P.L.F. - Um sistema adaptativo de detecção de intrusão em redes de computadores.

022/96 BRIGANTE, W.J.; MOREIRA, E.S. - Utilização de Monitores OLTP no gerenciamento de ambientes de manufatura integrada voltados à produção discreta.

021/95 BEZERRA, L.A.F.; SANTANA, R.H.C.; SANTANA, M.J. - Sistema auxiliar de arquivos baseado em disco WORM para ambientar computacional distribuído.

020/95 NUNES, M.G.V.; HASEGAWA, R. - PROTEMA: intelligent tutoring systems for mathematics.

019/95 OLIVEIRA, M.C.; TURINE, M.A.S.; MASIERO, P.C. - A statechart - based model for hypertext.

018/95 PIMENTEL, M.G.C. - Alternative operations for browsing hypertext.

017/94 ROMEIRO, N.M.L.; CASTELO FILHO, A. - Análise Comparativa de Métodos Numéricos de equações algebrico-diferenciais.

016/94 MAGALHÃES, A.L.C.C.; SIQUEIRA, M.F.; OLIVEIRA, M.C.F. - Operadores de Euler na modelagem por fronteira: conceito, aplicação, estudos de casos.