COPPE/UFRJ

METODOLOGIA DE ARMAZENAGEM E BUSCA DE DADOS SEMI-ESTRUTURADOS NA WEB

Carlos Augusto Sicsú Ayres do Nascimento

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro
Dezembro de 2009

METODOLOGIA DE ARMAZENAGEM E BUSCA DE DADOS

SEMI-ESTRUTURADOS NA WEB.

Carlos Augusto Sicsú Ayres do Nascimento

TESE SUBMETIDA AO CORPO DOCENTE DO INSTTUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc

_____

Prof. Alexandre Gonçalves Evsukoff, DR.

_____

Prof. Beatriz de Souza Leite Pires de Lima, D.Sc

_____

Prof. Antonio Cesar Ferreira Guimarães, D.Sc

_____

Prof. Elton Fernandes, D.Sc

RIO DE JANEIRO, RJ - BRASIL

DEZEMBRO DE 2009

Sicsú Ayres do Nascimento, Carlos Augusto

Metodologia de armazenagem e busca de dados semi-estruturados na Web / Carlos Augusto Sicsú Ayres do Nascimento - Rio de Janeiro: UFRJ/COPPE, 2009.

XV, 130 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2009.

Referências Bibliográficas: p. 125-128.

1. Mineração de Dados na Web. 2. Classificação por K-Vizinhos Mais Próximos. 3. Índice de Aproximação Estatítica. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

A Adrina e Matheus
A meus pais

AGRADECIMENTOS

A minha esposa Adriana e meu filho Matheus que por diversas vezes ficaram sem minha companhia em vários eventos.

Ao meu Orientador Nelson Ebecken, a quem devo tanto conhecimento adquirido ao longo destes anos e muito me incentivou durante os momentos mais difíceis.

Ao meu "irmão" José Luiz que sempre me incentivou e muito me ajudou na transformação de sonhos em realidade.

Aos meus amigos, Carlos Lemos, Renato Carr e Carlos Alexandre pelo apoio e ajuda, nos momentos difíceis.

Ao CNPQ pelo suporte financeiro que viabilizou a realização desta tese.

Ao Laboratório do Núcleo de Transferência de Tecnologia – NTT, pela infra-estrutura, suporte administrativo e logístico.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D. Sc.)

METODOLOGIA DE ARMAZENAGEM E BUSCA DE DADOS SEMI-ESTRUTURADOS NA WEB

Carlos Augusto Sicsú Ayres do Nascimento

Dezembro/2009

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Esta tese tem por objetivo desenvolver uma metodologia para automatizar o processo de busca, armazenamento e classificação de publicações científicas disponíveis na Internet, contribuindo para automatizar e facilitar a busca por documentos correlatos, além de propor um índice de aproximação estatística (IAE) para a reclassificação dos documentos reorganizando a lista de relevância produzida pela classificação por KNN. A metodologia proposta baseia-se em métodos de mineração de textos e contribui com as pesquisas voltadas a área de mineração de dados na Internet.

A metodologia foi desenvolvida em três módulos distintos e sua estrutura modular, permite a distribuição de processos em forma concorrente e/ou paralela para aumento do desempenho do processamento. Os resultados do estudo de caso são comentados para avaliar o desempenho da metodologia. Algumas conclusões e estudos futuros são apresentados.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D. Sc.)

METHODOLOGY OF STORAGE AND SEARCH OF SEMI-STRUCTURED DATA ON THE WEB

Carlos Augusto Sicsú Ayres do Nascimento

December/2009

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This thesis seeks the development of automated method to search, sort and store scientific publications available in the Internet. This automated method will contribute to facilitate the search of associated documents. It will also propose a Statistic Approximation Index (SAI) that will reprioritize the relevance list produced by the KNN categorization. The proposed process is based on data mining methodology, and it will contribute to Internet data mining research.

It includes three distinct modules allowing the distribution of process sequentially or in parallel, which will increase efficiency. The results of the case study are discussed to evaluate the performance of the methodology. Some conclusions and future studies are presented.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AG – Algoritmo Genético

DF – *Document Frequency*

e-Gov – Governo Eletrônico

HTTP - *HyperText Transfer Protocol*

IAE - Índice de Aproximação Estatística

IDFk - *Inverse Document Frequency*

KNN – K-*Nearest Neighbor*

PFP – Palavras Frequentes e Preditivas

RI – Recuperação da Informação

TS – Taxa de Similaridade

WWW - *World Wide Web*

UML - Linguagem de Modelagem Unificada (*Unified Modeling Language*)

Capítulo 1
Introdução

Desde as cavernas o homem transmite conhecimento por meio de dados armazenados nas paredes em forma de pictogramas [1]. Essas informações serviam para guiar novas gerações, demonstrando técnicas (como caçar corretamente, por exemplo) e alertas (presença de animais ou locais perigosos, e assim por diante), ou seja, serviam como comunicação.

Com o fim do nomadismo, o homem iniciou aglomerações em torno de feudos e reinos, até as cidades atuais. A invenção da escrita pelos egípcios, do papel pelos chineses, e da imprensa por Gutemberg [2], deu um grande impulso à troca de informação. Livros começaram a ser comercializados, difundindo idéias e conceitos. A informação não mais se perdia pelo tempo. Passaria a ser transmitida de geração a geração por meio da escrita. O homem passou a ter mais e mais informação disponível ao seu alcance.

Assim, a informação existia, mas havia dificuldade em organizá-la em conjunto, em obtê-la de acordo com sua necessidade. Dessa forma, nasceram as bibliotecas, onde os volumes podiam ser armazenados e resgatados de acordo com o interesse por dado assunto. Ao mesmo tempo, jornais e revistas traziam informações típicas da época de impressão, refletindo usos e costumes, além de permitir o estudo desta mesma época.

O avanço da ciência permitiu a invenção do computador eletrônico e seus meios de armazenamento permanente (discos rígidos, fitas magnéticas, CD's, DVD's, etc.). Da interligação dos computadores em redes cada vez maiores nasce a internet [3]. E com a internet, há o enorme salto na quantidade de informação manipulável, demandando técnicas de resgate de dados, permitindo que a informação seja obtida de forma organizada e rápida, sendo a base da sociedade da informação, segundo Wierner [4]. De acordo com ele, "... a soma da informação em um sistema é a medida de seu grau de organização; a entropia é a medida de seu grau de desorganização; um é o negativo do outro.". Ainda de acordo com Wierner: "A informação, as máquinas que a

tratam e as redes que ela tece são as únicas capazes de lutar contra essa tendência à entropia.”.

Ações de entidades civis e governamentais, como *e-Gov*, termo utilizado para governo eletrônico, disponibilizaram as pessoas tarefas que antes as obrigavam a se deslocar fisicamente até os locais onde essas tarefas eram oferecidas. Por exemplo, hoje em dia uma pessoa não precisa se deslocar fisicamente até a prefeitura para obter a segunda via do IPTU, ou ir a uma das sedes da Receita Federal para declarar seu imposto de renda. Basta que ela acesse a internet, e o serviço, se disponível, poderá ser executado. Compras em supermercados, pagamento de boletos bancários, saldos e extratos em conta corrente, são outros exemplos de serviços disponíveis pela grande rede.

Ao passo que os serviços passaram a ser oferecidos em maior quantidade, o uso de computadores passou a ser algo comum na vida do homem moderno. Em consequência, o armazenamento da informação em meios digitais tomou grandes proporções, muito em função de: preço de armazenamento muito baixo, quantidade de armazenamento muito alto, e velocidade de mudança das informações quase que instantânea, ao contrário dos livros em papel que uma vez impressos, não permitem correção da informação prestada incorretamente, necessitando de erratas anexadas às obras.

As empresas eram as grandes consumidoras do recurso “armazenamento de memória”. Usuários eram apenas entidades consultivas dessa informação armazenada. Não havia a figura do usuário doméstico. Porém, com passar do tempo, a criação e desenvolvimento da internet, entremeado com o *boom* da *web 2.0* e das redes sociais, usuários passaram a ser, também, os depositários de informação em grande quantidade na grande rede. Documentos, e não mais apenas dados tabulares, passaram a figurar nos sistemas de armazenamentos de *data centers*; currículos, anexos, relatórios dos mais diversos, necessitavam de índices que os remetiam ao assunto em debate no artigo armazenado.

A entrada em cena de outros usuários que não só os de informática, também trouxeram uma contribuição enorme à heterogeneidade de assuntos e formatos de dados. Lições de geografia, histórias de batalhas e dinastias, relatórios antropológicos, todos necessitavam da criação de índices de busca. Porém, o que se pode comprovar com a prática, é que esses usuários simplesmente armazenam suas informações em formato textual, dificultando em alguns casos, e impossibilitando em outros casos, o resgate do documento pela busca por assunto.

Assim, um número incalculável de documentos podem ser considerados correlatos, bastando, para tal, que o tema definido para a busca da informação armazenada seja amplo o suficiente.

As técnicas de busca por palavras chave não permitem, na maioria dos casos, o resgate da informação armazenada em sua totalidade. Isto decorre do fato de os textos possuírem quantificação semântica, não apenas um conjunto de dados alfanuméricos e simbólicos.

A busca dessas informações é tarefa árdua e difícil, pois as formas de pesquisa tradicionais trazem como resultado uma grande quantidade de respostas. Por exemplo, uma pesquisa realizada no *Google* [5] pela ocorrência "*text mining*" em 20/02/2007 produziu como resposta um conjunto com mais de quarenta milhões de resultados. O problema passa a ser o de se obter bons resultados e não apenas o de se obter resultados, o que torna necessária a criação de uma lista ordenada para determinar, de forma automatizada, os resultados a partir dos melhores resultados encontrados [6,7].

A automatização de um processo consiste na utilização de tecnologia, objetivando simplificar, minimizar ou eliminar a intervenção humana sobre esse processo, ou seja, torná-lo automático. A tendência constatada nos dias de hoje pode ser justificada pela necessidade de alta produtividade com um menor gasto de tempo. Para alcançar esse objetivo, a solução encontrada é a limitação da intervenção do homem no processo produtivo, ou seja, automatizar o processo de pesquisa. Por esse fato a automatização é tida como a essência das técnicas capitalistas de produção [8]. Como o objetivo deste trabalho de pesquisa é a busca por publicações, a intenção então passa a ser a de se obter uma lista ordenada com publicações relevantes de forma automatizada.

O contexto de um documento é formado pelos relacionamentos existentes (quando, onde e por quem foi escrito, além dos dados de publicação). Assim, um documento passa a exigir um contexto, pois um assunto pode não ter relevância em função do tempo de publicação. Documentos escritos em alguns locais podem ser mais relevantes do que documentos escritos em outros. Sendo assim, para algumas aplicações, pode ser muito importante que se analise o contexto junto com o documento.

De acordo com Tan [9], existe um potencial maior no uso de *Text Mining* do que no uso de *Data Mining*. "... 80% das informações de uma empresa estão armazenadas em documentos de texto".

Assim sendo, a busca destes documentos por assunto caminha na direção de técnicas que permitam a recuperação textual da informação, baseada em busca estatística dos dados, por meio de um argumento definidor de um tema. A variedade de conhecimento disponível na *web* em formato textual torna o ambiente propício para a aplicação de operações de mineração de textos.

1.1. Definição do Problema

Antigamente o conhecimento era disseminado através de revistas científicas, que eram distribuídas através de entregas postais. O tempo entre a publicação e a disseminação da informação era grande e permitia que a leitura pudesse ser feita aos poucos, ou seja, deveria ser feita até a próxima publicação. Com o advento da Internet, o tempo de disseminação da informação diminuiu muito, ocasionando que publicações ofertadas na Internet, possam ser consultadas em segundos, logo após a sua publicação. Atualmente existentes inúmeros locais que ofertam publicações e cada vez mais pesquisadores publicam seus trabalhos, fazendo com que tenhamos uma enorme base de conhecimento a nossa disposição. O tempo entre a publicação e a disseminação da informação não mais existe e a quantidade de publicações é enorme, tão grande que se tornou impossível ler todas as publicações disponíveis.

Como um pesquisador pode gerenciar suas pesquisas e trabalhos correlatos atualmente? Como ele pode obter informações e aproveitar-se delas, sem ler inúmeros trabalhos que não tem a menor relevância?

Uma metodologia capaz de efetuar pesquisas de forma automatizada, utilizando critérios definidos através de um perfil de interesse e feita sobre uma base de dados que centralize publicações relevantes ao assunto, pode permitir um melhor aproveitamento do tempo despendido para a leitura. Uma vez que a base pode ser alimentada com as principais publicações sobre os assuntos de interesse e será filtrada de forma automática, espera-se que uma lista ordenada por relevância gerada através do desenvolvimento de uma metodologia, possa atender a uma comunidade de pesquisadores.

Ao se utilizar desta metodologia, esta comunidade poderá aproveitar muito melhor o tempo gasto na finalidade de aquisição de novos conhecimentos.

Este trabalho terá como foco a busca estatística por um determinado tema em dados semi-estruturados. Essa busca se dará pela definição de um perfil de usuário, fornecido por meio de um texto que contenha o assunto correlato a encontrar e de alguns parâmetros de busca.

A metodologia de busca será apoiada pelo método KNN (K-*Nearest Neighbor*), ou K-Vizinhos mais Próximos [10], pela qual um indivíduo é classificado pela similaridade encontrada com K indivíduos previamente agrupados por coincidência de características, similaridade esta baseada no cálculo da distância Euclidiana:

$$d(x,y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2} \qquad \text{(Eq. 1)}$$

A análise dos resultados será feita à luz da comparação dos textos obtidos com o tema correlato utilizado como perfil. Assim sendo, um texto poderá ter todas ou parte das características do tema a ser encontrado.

Esta tese tem como objetivo a definição de uma metodologia de recuperação de informação oriundos da Internet. A metodologia define uma forma de busca e aquisição de publicações, preparação de dados e efetua uma classificação sobre grandes massas de dados através de tarefas automatizadas. São usadas técnicas de mineração de textos já consagradas, aplicadas em ambiente de computação de alto desempenho, com emprego de concorrência/paralelismo em máquinas com multiprocessadores. Uma nova concepção é apresentada visando á recuperação de informações com ganhos nos índices de Precisão e Recuperação de documentos.

1.2. Organização do Trabalho

O restante desta tese está organizado como se segue:

No capítulo 2 são apresentados conceitos teóricos envolvidos em mineração de textos.

No capítulo 3, discute-se o desenvolvimento da metodologia de busca, armazenamento e classificação de dados em um ambiente de sistema distribuído.

No capítulo 4, são apresentados os resultados obtidos pelo método proposto a partir de um estudo de caso com diferentes perfis.

No capítulo 5, são feitas considerações e observações sobre os resultados alcançados, além de considerações sobre trabalhos futuros.

Capítulo 2
Mineração de Textos

Mineração de textos (também conhecido como *Text Data Mining, Text Mining,* ou descoberta de conhecimento em bases de dados textuais) refere-se ao processo de extração de processos interessantes e não triviais de padrões a partir de documentos de texto com dados não estruturados [11]. Também é visto como uma extensão de Mineração de Dados (*Data Mining*), extração de conhecimento em bases de dados estruturados [9]. Atualmente com as tarefas de Mineração de Dados consolidadas, muitos esforços estão sendo aplicados à área de Mineração de Textos.

Diferentemente das técnicas de Mineração de Dados, as técnicas de Mineração de Textos visam efetuar buscas em dados não-estruturados ou semi-estruturados, e ainda não estão consolidadas, ou não são tão eficientes quanto às de Mineração de Dados. Mineração de Textos, entretanto, envolve tarefas muito mais complexas do que as tarefas de Mineração de Dados [9]. Muitas soluções foram propostas para agrupamento, sumarização e categorização de documentos, com o intuito de facilitar a recuperação da informação (RI).

Algumas ferramentas existentes oferecem suporte a extração de conhecimento em bases de dados não estruturados. Possuem suporte a diversas línguas entre elas, inglês, francês, holandês, alemão, italiano, português, espanhol dentre outras.

O *Text Mining for Clementine* [12], possibilita a extração de conceitos-chave, percepções e relação de dados. Faz a conversão de textos tais como documentos da suíte *Microsoft Office©*, anotações de *Call Centers*, respostas de pesquisas, emails, formulários da Internet e conteúdo de feeds RSS, para um formato capaz de ser processado por outras ferramentas das suítes de análises preditivas oferecidas pela SPSS para o gerenciamento e a execução de processos analíticos. O *Text Mining for Clementine* ainda permite a customização de dicionários de conceitos para áreas específicas de domínio. Pode ser usado para entender opiniões, atitudes e preferências

de clientes. Oferece ainda uma interface gráfica que se propõe a ser prática e fácil de usar.

O SAS *Text Miner* [13] também transforma os documentos de texto em um formato para facilitar a classificação, descoberta de associações e relacionamentos e o agrupamento de documentos. Faz conversão de formatos textuais, PDF©, HTML e documentos do *Microsoft Word*©. Executa a preparação de dados utilizando *stoplists* padrões que podem ser customizadas, executa tarefa de *stemming,* além de possuir dicionários para extração de partes específicas de informação. No SAS, os documentos são transformados em modelos no espaço n-dimensional e pode-se aplicar sobre eles algoritmos de agrupamento tais como: K-*means* ou SOM/Kohonen. Os termos podem ser organizados alfabeticamente, identificados em documentos, calculadas as freqüências e definidos pesos em funções de regras.

O *Statistica* [14] que aplica os conceitos de mineração de dados a informações não estruturadas, possibilitando a quantificação para extração de conhecimento através de sistemas de apoio a decisão.

Após o suporte a mineração de dados feitos pelos fabricantes de programas gerenciadores de bases de dados, os mesmos começam a oferecer tarefas de mineração de textos. O Oracle© [15], respeitado como o melhor gerenciador de banco de dados do mercado, na sua versão 10*g*, passou a oferecer suporte à mineração de textos. Entre as tarefas oferecidas, estão:

- Classificação por árvores de decisão;
- Agrupamento por K-means;
- Extração de recursos de forma interna sem exibição de resultados.

O SQL *Server*© [16], desenvolvido pela *Microsoft Corporation*©, vem tendo grande investimento em tecnologias para mineração de dados. Na versão 2005, por exemplo, ele possui nove algoritmos para esta tarefa, incluindo:

- Algoritmo para árvores de decisão e regressão;
- Agrupamento;
- Regressão linear;
- Redes neurais;

- Associação;
- Séries temporais, etc.

Para mineração de textos, o SQL Server converte textos não estruturados em uma estrutura de dados para análise através de relatórios, processamento analítico ou mineração de dados.

De acordo com PINHEIRO [17], das aplicações "puras" de Mineração de Textos, talvez o software comercial mais robusto e respeitado seja o TEMIS [18], principalmente pela grande quantidade de recursos para a criação de dicionários e tesauros que ele dispõe, fruto de intensas pesquisas desenvolvidas pela sua equipe de pesquisadores e técnicos. Já para softwares não comerciais, pode-se citar o TMSKJ e RIKText proposto por WEISS et al [18], o *Text Mine* proposto por KONCHADY em [18] e o que está em processo de desenvolvimento na COPPE/Civil, o Biguá, projeto que é um conjunto das pesquisas de vários estudantes de mestrado e doutorado deste centro, financiado pelo CNPq [18].

BASTOS [19], também apresenta ferramentas gratuitas e comerciais na área de mineração de textos, além de realizar trabalho significativo nesta área.

De acordo com REZENDE [11], a mineração de textos consiste das seguintes etapas:

a) Abordagem dos Dados:

- Análise Semântica: avaliam a sequência dos termos no contexto da frase, utilizando fundamentos e técnicas baseadas no processamento de linguagem natural;

- Análise Estatística: a importância dos termos é dada basicamente pelo número de vezes em que eles aparecem nos textos.

b) Preparação e Processamento dos Dados:

- Recuperação da Informação (RI): pode-se considerar esta etapa como sendo a primeira no processo de mineração de textos. Compreende o modelo booleano, onde um documento é representado por um conjunto de termos-índice, e o modelo de espaço vetorial, pelo qual os documentos são representados no espaço Euclidiano t-dimensional. Por este modelo, é criado um vetor contendo os termos e suas ocorrências respectivas, permitindo o cálculo da distância por meio de similaridade.

- Análise dos Dados: tem por objetivo a identificação de similaridade de significado entre as palavras, apesar de suas variações morfológicas.

A dificuldade em se analisar similaridade entre textos decorre do fato, por exemplo, de palavras diferentes poderem expressar a mesma ideia, apesar de suas diferenças morfológicas. Isto ocorre, por exemplo, no caso de palavras sinônimas. A noção de que sentenças podem assumir múltiplos significados conduz ao fenômeno das entradas distintas que devem assumir a mesma representação de significado [20,21]. Então, duas sentenças diferentes devem ter a mesma representação. Por exemplo [20]:

McDonalds tem comida vegetariana?

Eles têm refeição vegetariana no McDonalds?

Abaixo são descritas as etapas a serem cumpridas na mineração de um texto:

- ***Case Folding***

É um processo usado para acelerar as comparações a serem feitas no processo de preparação dos dados em *Text Mining*. Visa converter todos os caracteres a uma mesma forma de representação, todas maiúsculas ou todas minúsculas.

- ***Stopwords***

Um dos pontos importantes da preparação dos dados é a eliminação das palavras que não possuem representatividade sobre o texto. Essas palavras, normalmente utilizadas com alta incidência, não devem ser usadas para representar diferenças ou similaridades entre documentos, uma vez que são muito comuns. O conteúdo semântico dessas palavras é insignificante, e não são relevantes na análise dos documentos. Ao se retirar estas palavras do índice, reduzimos o espaço de armazenamento e melhoramos seu desempenho durante a tarefa de *Text Mining*. Uma lista básica de palavras pode ser usada para formar uma lista de stopwords, chamada de stoplist, de forma a melhorar a tarefa de *Text Mining*. Um conjunto de 138 palavras consideradas universais para a língua inglesa forma, então, uma *stoplist* a ser usada na preparação dos dados [22].

Uma tabela de *stoplist* com palavras universais para a língua inglesa pode ser vista no apêndice A, junto com uma tabela de termos usados na Web.

A maioria dessas palavras são conjunções, preposições ou pronomes que pouco contribuem para a formação do índice [22]. De acordo com a especialidade do texto, outras listas mais apropriadas podem ser criadas, mas, neste caso, os documentos a serem trabalhados são de gêneros distintos.

Outro conjunto de palavras (tais como: http, Web, etc) pode ser incluído à *stoplist* definida, já que levaremos em conta também documentos da Web. Estas palavras também pouco contribuem para a formação do índice.

- **Conflação**

É a tarefa de fusão ou combinação para igualar variantes morfológicas de textos. O principal método de conflação (*stemming*) é descrito a seguir [20,21,23].

- ***Stemming***

O processo de *Stemming* visa à redução ao radical (raiz), para que diferentes palavras com o mesmo significado não sejam contadas como palavras distintas. Normalmente, é usada uma palavra em uma consulta e as respostas podem trazer documentos não relevantes ao contexto exigido. Acontece que uma variante dessa palavra pode ser encontrada em outros documentos com maior relevância. Sendo assim, a consulta pode ser mais eficiente se trabalhar com os radicais das palavras ao invés da palavra original. Uma mesma palavra pode ter, dentre outras, variações como plurais, formas de gerúndio e acréscimo de sufixos [20,24,25]. A técnica consiste então em remover os sufixos e prefixos, para que palavras com o mesmo radical tenham significados similares [17,26].

O processo de *stemming* é dependente da linguagem, e as regras devem ser adaptadas para as diversas linguagens existentes, ou seja, os algoritmos devem ter suas regras modificadas de forma a atender a uma determinada linguagem. Para a língua inglesa podemos identificar as palavras *cats*, *catlike*, *catty* como tendo o radical *cat*. Outro exemplo pode ser visto nas palavras *stemmer*, *stemming* e *stemed* que possuem o radical *stem*. Assim, um algoritmo de *stemming* deve reduzir as palavras *fishing*, *fished*, *fish* e *fisher* à raiz *fish* [23].

A primeira publicação sobre o assunto surgiu em 1968 no trabalho sobre conflação de Julie Beth Lovins [27]. Em 1979 foi publicado o "stemming algorithm paper" [28], que foi a base para a criação do mais difundido algoritmo de *stemming*, o algoritmo de Porter [29]. Existem atualmente vários algoritmos de *stemming* disponíveis, alguns deles apresentados a seguir.

- ***Bruce Force Algorithm***

O termo que significa Algoritmo da Força Bruta, vem da concepção de pesquisas em inteligência artificial e a forma de resolver problemas denominada força bruta. Este método consiste na criação de uma tabela de relacionamento entre a raiz das palavras e as formas de inflexão. Se uma inflexão for encontrada, ela será substituída pelo radical associado. Este método é criticado pelo fato de outros métodos convergirem mais facilmente, e pela necessidade de armazenamento de uma grande lista [23].

- ***Stemming S Algorithm***

Este método é o mais simples, mas, mesmo assim, é bastante usado por ser conservador e prático. Ele se baseia na transformação das palavras no plural para a sua forma singular. A tabela 2.1 contém as regras básicas do algoritmo para a língua inglesa.

Tabela 2.1 – Regras básicas do algoritmo de *stemming S* [30].

Regras
Se a palavra termina em “ies”, mas não em “eies” ou em “aies” Então substituir “ies” por “y” Fim-se;
Se a palavra termina em “es”, mas não em “aes” ou “ees” ou “oes” Então substituir “es” por “e” Fim-se;
Se a palavra termina em “s”, mas não em “us” ou em “ss” Então substituir “s” por “” (nulo ou vazio) Fim-se;

- ***Suffix Stripping Algorithm***

Este método não verifica uma tabela. Ele atua na inflexão e no relacionamento com o radical, onde é entrada a palavra e a mesma é retornada em forma de raiz. Trabalha com um conjunto de regras, conforme os exemplos da tabela 2.2, sendo mais simples de implantar que o método da força bruta.

Tabela 2.2 – Alguns exemplos de regras do *Suffix Stripping Algorithm*.

Regra
Se a palavra termina em “*ed*” Então remover “*ed*” da palavra Fim-se;
Se a palavra termina em “*ing*” Então remover “*ing*” da palavra Fim-se;
Se a palavra termina em “*ly*” Então remover “*ly*” da palavra Fim-se;

- ***Stochastic Algorithm* – Algoritmo Estocástico**

É um método probabilístico para identificar o radical de uma palavra, por meio de uma tabela de relacionamento entre o radical e a inflexão. Este modelo é normalmente expressado em uma forma complexa de regras linguísticas, similar à natural, para retirada de sufixos [23].

- ***Porter Stemming Algorithm***

É também um método de eliminação automática de sufixos, para reduzir um termo a uma forma radical para palavras com mais de duas letras. É usado para tratar um conjunto de aproximadamente 60 sufixos através de eliminações e transformações sucessivas da palavra original, sendo um método bastante consagrado. Possui versões em diversas linguagens de programação, tais como: ANSI C, Java, Perl, Python, C Sharp, Visual Basic, dentre outras. Existem ainda versões para Prolog, Haskell, T-SQL e MathLab. Seu método consiste em um conjunto de cinco passos a serem usados em ordem para eliminação de sufixos [29,31].

Segue uma breve descrição do primeiro passo, o qual é dividido em três partes, e usado para tratar plurais e particípios passados. Na primeira parte são tratados os sufixos de plural, tais como os finais de palavras “sses”, “ies”, “ss” e “s”. Por exemplo, a palavra “caresses” é transformada na palavra “caress”; a palavra “ponies” é transformada em “poni”. Na segunda parte são eliminados os sufixos “eed”, “ed” e “ing”, de acordo com o tamanho da palavra e, em seguida, são feitas correções para completar a palavra para a sua forma usual. Como exemplo, tem-se a palavra “agreed” que é transformada em “agree”, ou a palavra “conflated”, que primeiro é transformada em “conflat” e posteriormente é corrigida para “conflate”. O mesmo ocorre com a palavra “filing”, que primeiro é transformada em “fil” para posteriormente ser transforma em “file”. Na terceira parte, palavras que tem um determinado sufixo são verificadas para se determinar se devem ter esses sufixos alterados, tais como “happy” para “happi” e “sky”, que se mantém como “sky”.

- **Método de Lovins**

Este método trabalha com apenas um passo para a eliminação automática de aproximadamente 250 sufixos. Ele começa eliminando o sufixo mais longo até chegar aos sufixos mais curtos. Em um teste com um documento de 8460 palavras, ele reduziu em 39% a lista de palavras, enquanto o método de Porter reduziu o mesmo documento em apenas 29%. Por exemplo, a palavra "heating" pelo método de Porter é reduzida ao radical "heated", e pelo método de Lovins é reduzida ao radical "heat", que é a forma correta de redução da palavra "heating" [27,32].

2.1. Recuperação da Informação

Segundo Ferneda [33] o termo "Recuperação da Informação" significa para alguns pesquisadores: "é a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário". Para outros: "consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondente". Ferneda define ainda que "o termo pode ser empregado para designar a operação que fornece uma resposta mais ou menos elaborada a uma demanda, e esta resposta é convertida num produto cujo formato é acordado com o usuário".

Este trabalho de pesquisa se utiliza de processos de busca de informação na Internet, para adquirir resumos de documentos referentes a publicações científicas, sua extração de dados e posterior recuperação, de acordo com um perfil definido pelo usuário. O contexto da pesquisa forma o perfil do usuário e o mesmo é formado por resumo de um artigo de seu interesse, a definição dos termos mais relevantes e seus respectivos pesos para aproximação ao domínio do assunto. O usuário terá ainda a possibilidade de utilizar regras de consulta para filtragem dos dados.

Um documento então é definido pela sua representação (conjunto de palavras relevantes) e é recuperado através de uma função de busca e opcionalmente por uma expressão de busca. A função de busca aqui usada é a classificação por *KNN* e a expressão de busca será definida para a filtragem dos resultados de acordo com as necessidades do usuário.

2.2. Métricas de Avaliação de Resultados

Uma vez executados os passos relativos à extração e à análise dos dados, dá-se o efetivo processo de recuperação da informação, o que demanda a definição de um sistema de métrica para que se possa avaliar os resultados obtidos (quantos documentos recuperados estão contidos na categoria desejada).

De REZENDE [11] obtém-se que as métricas podem utilizar quatro abordagens:

- Frequência de Documentos: ou *Document Frequency (DF)*, na qual a remoção de termos raros é efetuada, por serem, os termos, não-informativos. Esta técnica tem como efeito colateral a redução da dimensionalidade do espaço de características;

- Informação Mútua: a qual leva em conta a quantidade *A* de vezes em que um termo *t* e uma categoria *c*, co-ocorrem. É um critério normalmente usado em modelagem estatística da linguagem em associações de palavras e aplicações correlatas;

- Estatística χ^2: a qual mede a falta de independência entre um termo *t* e uma categoria *c*, e pode ser comparada à distribuição χ^2 com um grau de liberdade para julgar extremos.

- Ganho de Informação: a qual mede o número de partes de informação obtidas para predição da categoria, pela presença ou ausência de um termo em um documento. Por esta técnica, termos cujos ganhos de informação forem menores que um determinado valor são retirados do espaço de busca considerado. Com isso, a estimativa de complexidade de tempo é *O(N)* e a complexidade de espaço é *O(V,N)*, onde *N* é o número de documentos e *V* é o tamanho do vocabulário. A quantidade de informação esperada para determinar a classificação de uma dada amostra é apresentada em SILVA [32],

$$I(S1, S2, ..., Sm) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \text{(Eq. 2)}$$

Onde:

I (S1, S2, ..., Sm) : informação necessária para classificar um documento numa dada categoria (ou,como relacionado neste trabalho, num dado perfil).

p_i : probabilidade de um documento pertencer à *i-ésima* categoria, obtida como,

$$p_i = \frac{s_i}{S} \qquad \text{(Eq. 3)}$$

Sendo:

s_i : número total de amostras na categoria *i*.

S : número total de amostras.

A entropia (que mede o grau de aleatoriedade dos valores dos termos a serem utilizados na busca) é determinada como [34]:

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + ... + S_{mj}}{S} I(S_{1j}, ..., S_{mj}) \qquad \text{(Eq. 4)}$$

Onde:

E (A) : é a entropia com a partição do atributo *A* ou devido ao atributo *A*.

m : representa o número de categorias em que um dado documento pode ser classificado (de acordo com a definição do perfil por parte do usuário).

$j = 1, ..., v$: representa o número de categorias de uma dada variável preditora (neste caso, de um dado termo utilizado na busca).

$(S_{1j} + ... + S_{mj}) / S$: peso correspondente ao número de amostras na categoria *j* da variável preditora, dividido pelo total de amostras.

$I(S_{1j}, ..., S_{mj})$: informação esperada para a categoria *j* da variável preditora.

Ainda de acordo com SILVA [32], o ganho da informação é a redução esperada da entropia, devendo-se calcular para cada atributo, o seu ganho (neste trabalho, os atributos são os termos utilizados na busca, ao passo que as classes são as categorias desejadas), determinada por Han & Kamber [34] como:

$$Gain\ (A) = I\ (S_1\ ,\ S_{2\ ,}\ \ldots\ ,\ S_m) - E\ (A) \qquad \text{(Eq. 5)}$$

Assim, da Equação 5 conclui-se que o objetivo primeiro para o ganho da informação é a diminuição do grau de entropia ao seu valor mínimo: 0 (zero).

2.3. Índice de Aproximação Estatística (IAE)

De forma a garantir o máximo de rendimento na recuperação de documentos com assuntos correlatos, propõe-se utilizar um novo índice, denominado **Índice de Aproximação Estatística, ou IAE**. Este índice é utilizado em uma segunda ordenação dos documentos (previamente ordenados em função de suas similaridades - distâncias - para o documento perfil), atuando no sentido de reduzir a entropia ao máximo. Desse modo, a equação alterada ficaria como,

$$Gain\ (A) = I\ (S_1\ ,\ S_{2\ ,}\ \ldots\ ,\ S_m) - E\ (A)\ +\ IAE\ (A) \qquad \text{(Eq. 6)}$$

O valor de **IAE** é obtido pelo emprego de *n* termos de grande relevância para o assunto considerado. Há de se observar aqui, que o uso deste índice difere da proposição feita por Salton e McGill, e descrita em Zuchini [35], para o uso do *IDFk* (*Inverse Document Frequency*), que é o inverso da freqüência de ocorrência do termo *k* em relação ao total de documentos. Na proposição apresentada aqui, o índice **IAE** será obtido ponderando-se cada termo a ser utilizado na categorização do documento, somando-se cada contribuição individual para compor o novo índice. Por exemplo, se forem utilizados *n* termos: $t_1,t_2,t_{3,\ldots,}t_n$ e, supondo-se que cada termo tem como grau de importância para o assunto considerado os seguintes pesos: $w_1,w_2,w_{3,\ldots,}w_n$,

respectivamente, e supondo-se ainda que a ocorrência destes termos em valores absolutos no documento *i* seja $o_{1i}, o_{2i}, o_{3i}, \ldots, o_{ni}$, então, o Índice de Aproximação Estatística para esse documento seria calculado como,

$$IAE_i = \frac{\sum_{j=1}^{n} w_i \times o_{ji}}{\sum_{j=1}^{n} o_{ji}} \qquad \text{(Eq. 7)}$$

Este índice, uma vez obtido, é listado decrescentemente, fornecendo a ordem dos documentos por grau de proximidade ao assunto considerado.

Em MATSUNAGA [36], discute-se a atribuição de pesos aos termos, em função da importância dos mesmos. POPESCUL e UNGAR [37] definem uma forma prática e eficiente para a definição dos termos mais relevantes, denominada "Palavras Freqüentes e Preditivas", baseada no produto:

$$p(\text{word} \mid \text{class}) \; X \; \frac{p(\text{word} \mid \text{class})}{p(\text{word})} \qquad \text{(Eq. 8)}$$

Esta abordagem leva em conta a freqüência e preditividade de palavras a partir da ocorrência das mesmas em um dado assunto (classe) e suas respectivas ocorrências em todo o domínio de assuntos. É calculado pelo quadrado da ocorrência de uma determinada palavra em um determinado assunto dividido pela ocorrência desta palavra em todo o domínio

Diferentemente do uso discutido por eles, onde se pretende nomear agrupamentos de documentos, o seu uso como forma de identificar as palavras mais relevantes é indicada para se definir os termos a serem utilizados pelo índice IAE, uma vez que se baseia na freqüência de palavras por assunto sobre a freqüência destas palavras sobre todos os assuntos. Para o caso de uma mesma palavra apresentar freqüência em mais de um assunto, a mesma será substituída em ambos os assuntos pela próxima palavra da lista de maior relevância, evitando-se assim ambigüidades entre diferentes assuntos.

2.4. Comparação de Desempenho

De Liu [32], obtém-se que os dois principais índices utilizados para comparar o desempenho na recuperação são: a **precisão / *precision*** e a **recuperação** / ***recall***.

A "precisão" é a quantidade de documentos relevantes recuperados em relação ao total de documentos recuperados.

A "recuperação" é a quantidade de documentos relevantes recuperados em relação ao total de documentos relevantes.

Suponha que a coleção a ser considerada contenha 20 documentos. Fornecido um perfil, sabe-se que 8 documentos são relevantes.

Para exemplificar, Liu [38] constrói uma tabela considerando uma coleção com 20 documentos e 8 documentos resultantes, onde a coluna "Relevância (+/-)" informa se um dado documento é ou não relevante para o contexto considerado:

Tabela 2.3: Valores de Precisão e Recuperação

Documento nº	**Relevância (+/-)**	**Precisão**	**Recuperação**
1	+	1/1 = 100%	1/8 = 13%
2	+	2/2 = 100%	2/8 = 25%
3	+	3/3 = 100%	3/8 = 38%
4	-	3/4 = 75%	3/8 = 38%
5	+	4/5 = 80%	4/8 = 50%
6	-	4/6 = 67%	4/8 = 50%
7	+	5/7 = 71%	5/8 = 63%
8	-	5/8 = 63%	5/8 = 63%
9	+	6/9 = 67%	6/8 = 75%
10	+	7/10 = 70%	7/8 = 88%
11	-	7/11 = 63%	7/8 = 88%
12	-	7/12 = 58%	7/8 = 88%
13	+	8/13 = 62%	8/8 = 100%
14	-	8/14 = 57%	8/8 = 100%
15	-	8/15 = 53%	8/8 = 100%
16	-	8/16 = 50%	8/8 = 100%
17	-	8/17 = 53%	8/8 = 100%
18	-	8/18 = 44%	8/8 = 100%
19	-	8/19 = 42%	8/8 = 100%
20	-	8/20 = 40%	8/8 = 100%

De posse dos dados da Tabela 2.3, constrói-se a curva precisão-recuperação (*precision-recall curve*), conforme a figura 2.1:

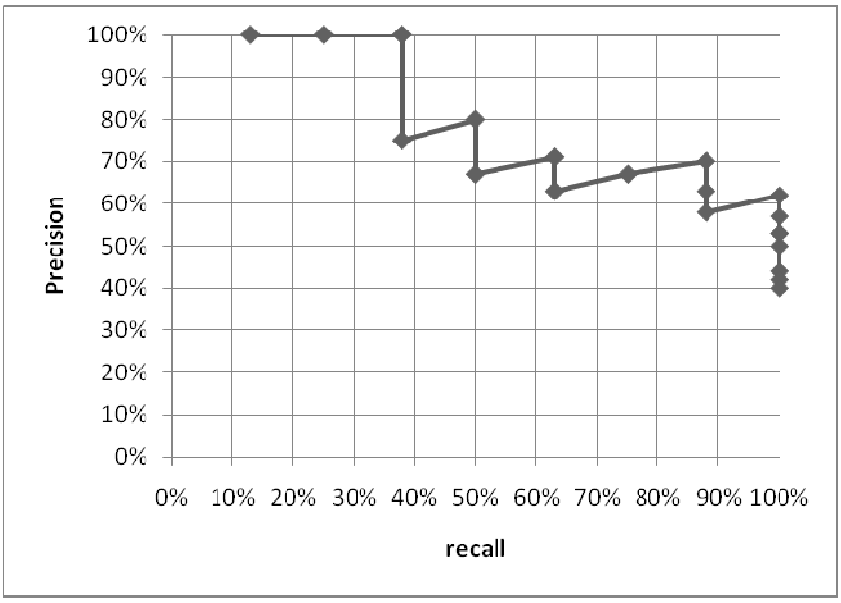


Figura 2.1: Curva Precisão-Recuperação

Essa curva é bastante útil quando se deseja comparar diferentes resultados de recuperação. Por exemplo, no gráfico da figura 2 (obtido de Liu [38]) podem ser vistos os resultados de dois testes hipotéticos.

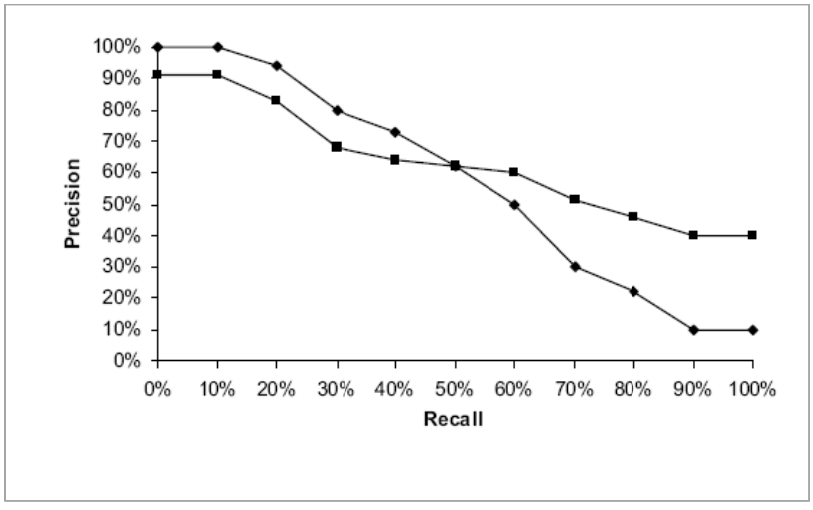


Figura 2.2: Curvas Precisão-Recuperação Comparativas

A análise das duas curvas permite observar que um dos métodos tem melhor precisão que o outro quando o índice de recuperação (*recall*) é baixo, enquanto que a precisão cai à medida que este índice sobe.

Capítulo 3
Descrição da Metodologia

3.1. Metodologias de Busca e Aquisição, Preparação de Dados e de Pesquisa

A Internet é um conglomerado infindável de redes eletrônicas interconectadas, criando um meio global de comunicação. Atualmente, a Internet é uma das áreas que mais cresce na informática. Com a atual expansão, já temos todo o globo terrestre interligado, trocando informações, fazendo negócios e interagindo entre si.

Existe uma grande quantidade de serviços disponíveis na Internet, entre eles o mais popular é o WWW (*World Wide Web*), devido a uma interface gráfica interativa, simples e de fácil utilização, em que qualquer usuário desfruta dos seus recursos. A WWW baseia-se no conceito de *HiperTexto*, onde palavras e imagens possuem referências próprias ou *links* com outros pontos no mundo, facilitando assim a navegação pela extensa rede.

Os portais hoje disponíveis na internet possuem centenas de milhões de publicações científicas distribuídas em centenas de diferentes sites. Cada um destes sites armazena milhões de publicações organizadas sobre diferentes assuntos, no passado isso era suficiente para se efetuar buscas, mas hoje em dia a multidisciplinaridade faz parte da realidade na publicação de um artigo científico e força a uma integração entre diferentes pesquisadores. Isso faz com que apenas organizar essas publicações por assunto impeça que alguns artigos importantes correlacionados possam não ser encontrados. Um armazenamento único com publicações obtidas dos sites mais relevantes pode permitir uma forma de busca mais eficaz.

A aquisição destes artigos então passa a ser fundamental para o sucesso deste projeto de pesquisa. Não foram localizadas ferramentas capazes de efetuar esta aquisição, sendo necessário que uma metodologia para esta tarefa fosse aqui proposta. Para a extração e preparação dos dados, é necessária uma metodologia capaz de extrair

os dados de diferentes estruturas de armazenamento semi-estruturado. Esta metodologia deve ser capaz de extrair dados a partir de diferentes formatos de estruturação. Para esta tarefa também não foram encontradas ferramentas disponíveis, o que obrigou a definição de uma forma de extração que fosse adaptável aos diferentes sites de publicação.

A metodologia se divide em três diferentes tarefas, sendo elas a busca e aquisição de dados na internet feita através do Módulo Buscador, a preparação e armazenamento de dados em formato estruturado feita pelo Módulo Preparador e, por fim, a classificação de documentos de maior interesse feita pelo Módulo Pesquisador.

Os módulos se dividem de acordo com a metodologia empregada:

1. O Módulo Buscador utiliza-se da metodologia para Busca e Aquisição de Dados e é descrito na seção 3.2 deste capítulo;
2. O Módulo Preparador utiliza-se da metodologia para Preparação e Armazenamento de Dados e é descrito na seção 3.3 deste capítulo;
3. O Módulo Pesquisador utiliza-se da metodologia para Pesquisa na Base de Armazenamento de Dados Estruturados e é descrito na seção 3.4 deste capítulo.

A arquitetura da ferramenta desenvolvida se divide em diferentes tarefas que podem ser distribuídas em diferentes máquinas. O servidor principal tem por função manter a base de dados do sistema e atender as requisições de dados das diferentes tarefas. O usuário credenciado para o uso do sistema define o seu perfil que será utilizado na pesquisa, seja através da inclusão do resumo de uma tese ou de um artigo de seu interesse. Este perfil é, então, utilizado para a classificação das publicações disponíveis no servidor, que é feita sobre os resumos das publicações armazenadas. Após a classificação, é criada uma lista ordenada com as publicações mais relevantes. O sistema se incumbirá de efetuar a classificação e de gerar a lista, além de atualizar a lista sempre que se fizer necessário, de forma automática após a inclusão de novas publicações. Com o intuito de aumentar a eficiência, foi desenvolvida uma metodologia de reclassificação dos resultados, levando-se em conta um grau de relevância determinado pelo Índice de Aproximação Estatística (IAE), o qual contribui para um ganho de precisão no emprego da metodologia.

3.2. O Módulo Buscador

Um dos grandes problemas é a busca de documentos através da internet. Estes documentos geralmente estão armazenados em diferentes locais e necessitam de muito tempo de busca e aquisição. O Módulo Buscador irá acessar de forma automatizada determinados locais, para a busca de publicações e fará a aquisição das páginas de conteúdo para o armazenamento. Esta parte da metodologia permite que um navegador especial acesse a internet e busque em determinados sites os arquivos de conteúdo ali armazenados para a posterior extração e armazenamento estruturado dos dados feito pelo Módulo Preparador. Estes sites armazenam páginas contendo informações das publicações e seus respectivos resumos. Ao fazer acesso a esses locais, passa a ser possível trazer o conteúdo de cada um destes documentos, o qual não precisará ser lido e o conteúdo obtido na página será armazenado na máquina local na forma de arquivo de computador. O grande problema está no fato de que as páginas de conteúdo não possam ser acessadas diretamente de forma sequencial, mas um conjunto de páginas de opções contendo *links* para as páginas de conteúdo, podem ser acessadas de forma sequencial. A metodologia faz uso de dois diferentes modos de contexto para a realização desta tarefa, sendo eles o Modo de Busca de *Links* e o Modo de Aquisição e Armazenamento, que são descritos da seção 3.2.1 e 3.2.2 respectivamente. A figura 3.1 apresenta o algoritmo que realiza a tarefa de Busca e Aquisição, mudando o seu contexto após a aquisição dos *links* no modo de Busca e passando ao modo de Aquisição e Armazenamento e retornando ao modo de Busca, até que o total de publicações definida pelo operador seja atingido.

```
Para a quantidade de publicações definidas por assunto escolhido
    Para cada página de opções
        Obter os links das publicações (Modo Busca de Links)
        Para cada um dos dez links obtidos (Modo Aquisição e Armazenamento)
            Obter a publicação e gravar o arquivo em diretório próprio na
                máquina local
        Fim-Para
    Fim-Para
Fim-Para
```

Figura 3.1 – Algoritmo para aquisição e armazenamento de dados.

3.2.1 O Modo de Busca de *Links*

No primeiro modo de contexto, para cada página de opções é possível obter os *links* apropriados respectivos a cada publicação, sendo que apenas um grupo de dez *links* são apresentados por vez em cada página. Isso ocorre com a maioria dos sites visitados durante o desenvolvimento deste trabalho, mas seu valor é parametrizado, permitindo que caso este valor mude, o mesmo pode ser facilmente redefinido. As páginas de opções, estas sim são apresentadas de forma seqüencial com variação numérica de valores de múltiplos de dez, sendo então possível acessar qualquer página de opções de forma sequencial ou aleatória. A metodologia aqui apresentada é capaz de obter estas páginas de opções e definir uma lista *links* que é usada para acessar e obter as páginas de conteúdo. Com esta metodologia, as páginas de conteúdo referentes a cada publicação podem então ser adquiridas da Internet de forma sequencial e/ou aleatória automaticamente. As páginas de opções são obtidas de forma sequencial a partir da alteração de forma sequencial da página em intervalos de dez publicações e cada link pode ser montado. A figura 3.2 apresenta uma página de opções e a figura 3.3 exibe uma parte de seu código fonte onde cada link é obtido.

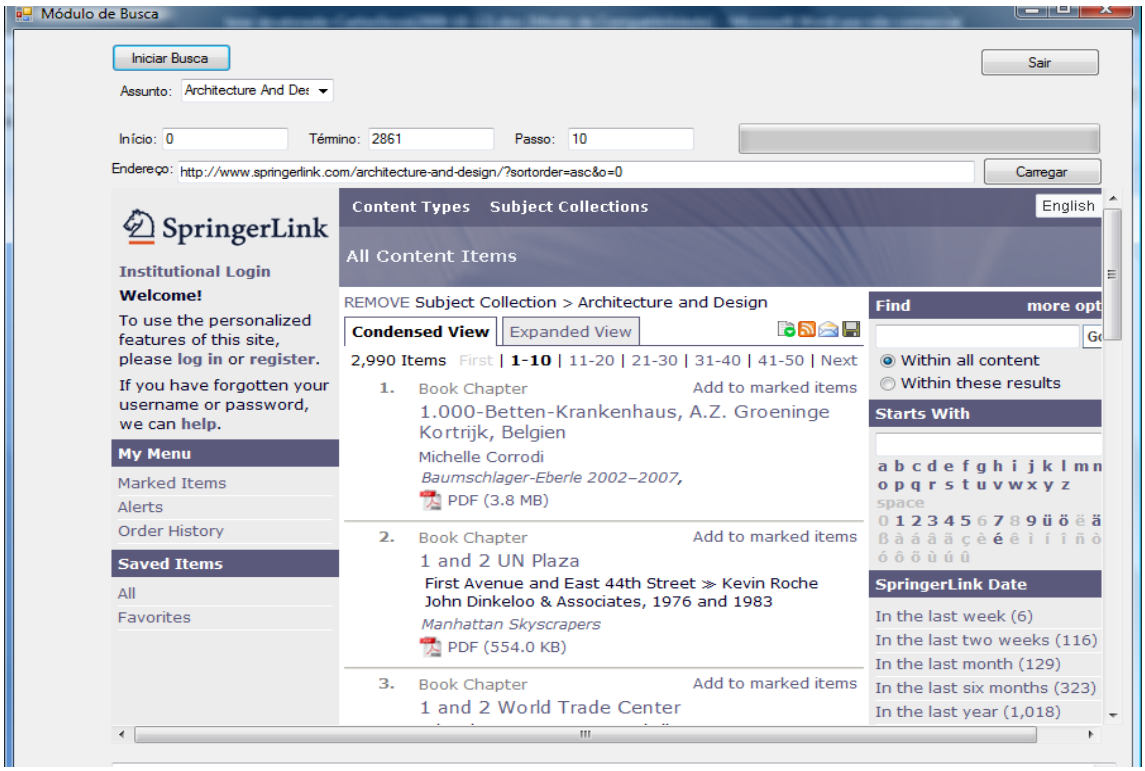
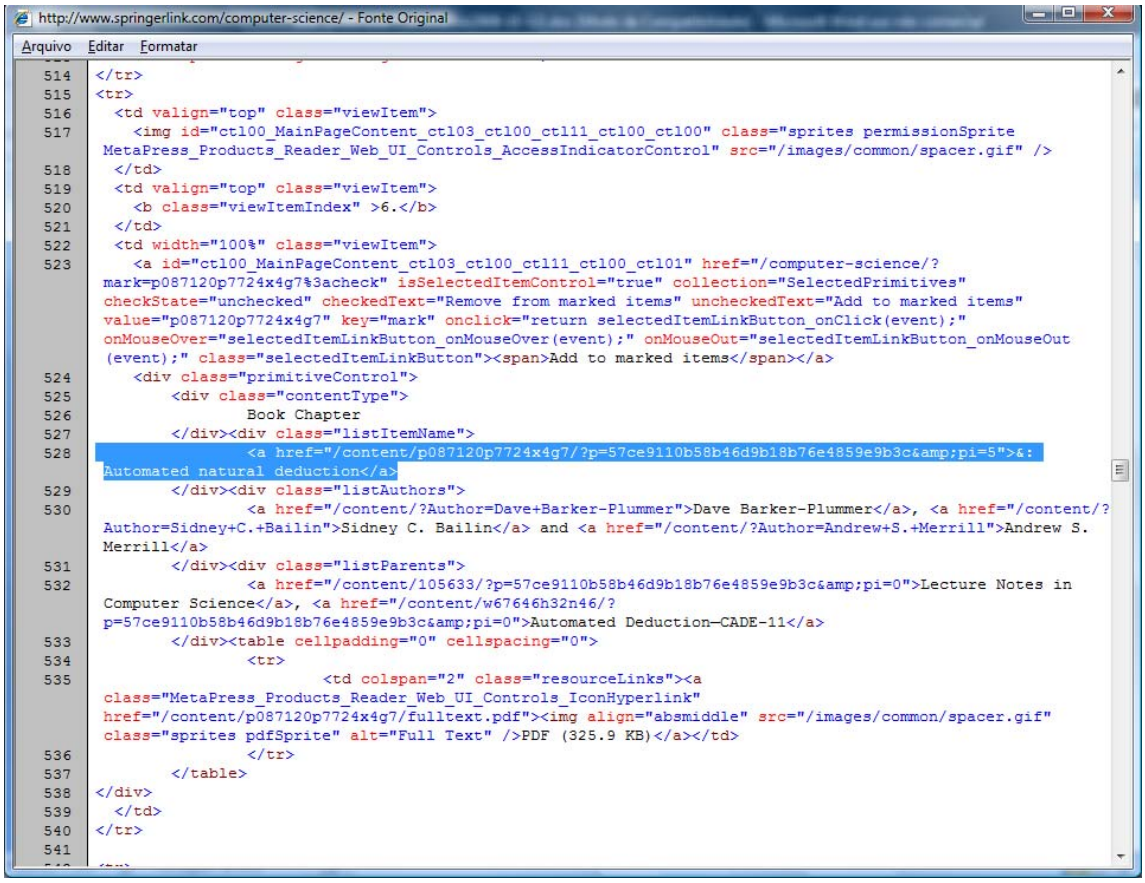


Figura 3.2 – Página de opções.




```
514 </tr>
515 <tr>
516   <td valign="top" class="viewItem">
517     <img id="ctl00_MainPageContent_ctl03_ctl00_ctl11_ctl00_ctl00" class="sprites permissionSprite
    MetaPress_Products_Reader_Web_UI_Controls_AccessIndicatorControl" src="/images/common/spacer.gif" />
518   </td>
519   <td valign="top" class="viewItem">
520     <b class="viewItemIndex" >6.</b>
521   </td>
522   <td width="100%" class="viewItem">
523     <a id="ctl00_MainPageContent_ctl03_ctl00_ctl11_ctl00_ctl01" href="/computer-science/?
    mark=p087120p7724x4g7%3acheck" isSelectedItemControl="true" collection="SelectedPrimitives"
    checkState="unchecked" checkedText="Remove from marked items" uncheckedText="Add to marked items"
    value="p087120p7724x4g7" key="mark" onclick="return selectedItemLinkButton_onClick(event);"
    onMouseOver="selectedItemLinkButton_onMouseOver(event);" onMouseOut="selectedItemLinkButton_onMouseOut
    (event);" class="selectedItemLinkButton"><span>Add to marked items</span></a>
524     <div class="primitiveControl">
525         <div class="contentType">
526                 Book Chapter
527         </div><div class="listItemName">
528                 <a href="/content/p087120p7724x4g7/?p=57ce9110b58b46d9b18b76e4859e9b3c&amp;pi=5">&:
    Automated natural deduction</a>
529         </div><div class="listAuthors">
530                 <a href="/content/?Author=Dave+Barker-Plummer">Dave Barker-Plummer</a>, <a href="/content/?
    Author=Sidney+C.+Bailin">Sidney C. Bailin</a> and <a href="/content/?Author=Andrew+S.+Merrill">Andrew S.
    Merrill</a>
531         </div><div class="listParents">
532                 <a href="/content/105633/?p=57ce9110b58b46d9b18b76e4859e9b3c&amp;pi=0">Lecture Notes in
    Computer Science</a>, <a href="/content/w67646h32n46/?
    p=57ce9110b58b46d9b18b76e4859e9b3c&amp;pi=0">Automated Deduction—CADE-11</a>
533         </div><table cellpadding="0" cellspacing="0">
534                 <tr>
535                         <td colspan="2" class="resourceLinks"><a
    class="MetaPress_Products_Reader_Web_UI_Controls_IconHyperlink"
    href="/content/p087120p7724x4g7/fulltext.pdf"><img align="absmiddle" src="/images/common/spacer.gif"
    class="sprites pdfSprite" alt="Full Text" />PDF (325.9 KB)</a></td>
536                 </tr>
537         </table>
538 </div>
539   </td>
540 </tr>
541
```

Figura 3.3 – Código fonte da página de opções.

Um *link* para as páginas de opções montado é mostrado através da figura 3.4, onde são definidos o assunto da publicação e o contador de páginas.

```
URL = "Site Principal" + assuntoPublicacao + FormaDeBusca + contadorPaginas;
```

Figura 3.4 – *Link* montado para aquisição de uma página de opções.

A URL é montada como a seguir:

1. Inicia-se pelo site principal, como por exemplo: "HTTP://ww.springerlink.com";
2. Acrescenta-se o tipo de publicação (assunto), que se encontra em uma variável. Seu conteúdo é escolhido pelo usuário antes do início da busca e aquisição, tal como: "*computer-science*";
3. Inclui-se a informação de forma de busca específica ao site escolhido, que pode ser: "/?sortorder=asc&o=";

4. E ao final é incluído o valor da sequência em múltiplo da quantidade de links por página de conteúdo. Este valor é definido pelo valor inicial de busca configurado pelo usuário e é incrementado pelo valor do parâmetro passado. Exemplo: 10 unidades.

A ferramenta então em modo de aquisição de links varia apenas o valor final do link montado de acordo com o assunto. A figura 3.5 apresenta uma sequência de *links* para as páginas de opções que serão acessadas para aquisição através dos *links* das páginas de conteúdo das publicações pelo assunto definido. Pode-se observar que existe apenas a variação do valor final em cada link, podendo então as páginas de opções serem acessadas de modo sequencial ou aleatório.

http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **0**
http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **10**
http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **20**
.
.
.
http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **2840**
http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **2850**
http://www.springerlink.com/architecture-and-design/?sortorder=asc&o= **2860**

Figura 3.5 – Sequência de links que serão usados para aquisição das páginas de opções para o assunto definido.

3.2.2 O Modo de Aquisição e Armazenamento

No modo de aquisição e armazenamento, o processo se comporta de maneira diferente e uma vez que já se encontra de posse dos *links* para as páginas de conteúdo, o contexto passa para aquisição dos arquivos das publicações. Na figura 3.6 pode-se observar a formação dos links adquiridos, nela pode-se verificar que os *links* tem um mesmo formato, mas os conteúdos destacados em negrito são muito diferentes e esta diferença não permite a busca seqüencial diretamente sobre as publicações e obriga a

execução em modo de busca. A ferramenta então no modo de Aquisição e Armazenamento pode buscar cada publicação independentemente das demais, carregá-la e armazená-la em um diretório local na máquina usada para esta tarefa.

<a href="/content/**101560**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **0**">
<a href="/content/**156a15qjxchla0bd**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **1**">
<a href="/content/**hte0dbqu8ampyet7**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **2**">
<a href="/content/**295lw57lqt4ulwq1**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **3**">
<a href="/content/**wf57lpx6ltayengq**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **4**">
<a href="/content/**p087120p7724x4g7**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **5**">
<a href="/content/**dm682m4186j63481**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **6**">
<a href="/content/**77n72w18103g5708**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **7**">
<a href="/content/**f887586xj824803j**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **8**">
<a href="/content/**f778450250908wg7**/?p=57ce9110b58b46d9b18b76e4859e9b3c&pi= **9**">

Figura 3.6 – Links adquiridos da página de opções.

A figura 3.7 apresenta uma página de conteúdo com uma publicação sendo adquirida, a seguir o arquivo obtido é armazenado no diretório definido na máquina local. Para cada diferente assunto é definido um diretório próprio para o armazenamento. Para evitar perda de desempenho em função de um número muito grande de arquivos que um determinado assunto possa ter, os mesmos são depositados em diretórios numerados e cada diretório armazena no máximo 1.000 arquivos. Isto permite que diferentes instâncias sejam executadas de forma independente diminuindo o tempo de busca e aquisição. A execução se mantém neste contexto até que todas as páginas de conteúdo sejam armazenadas e ao término, o contexto retorna ao modo de Busca.

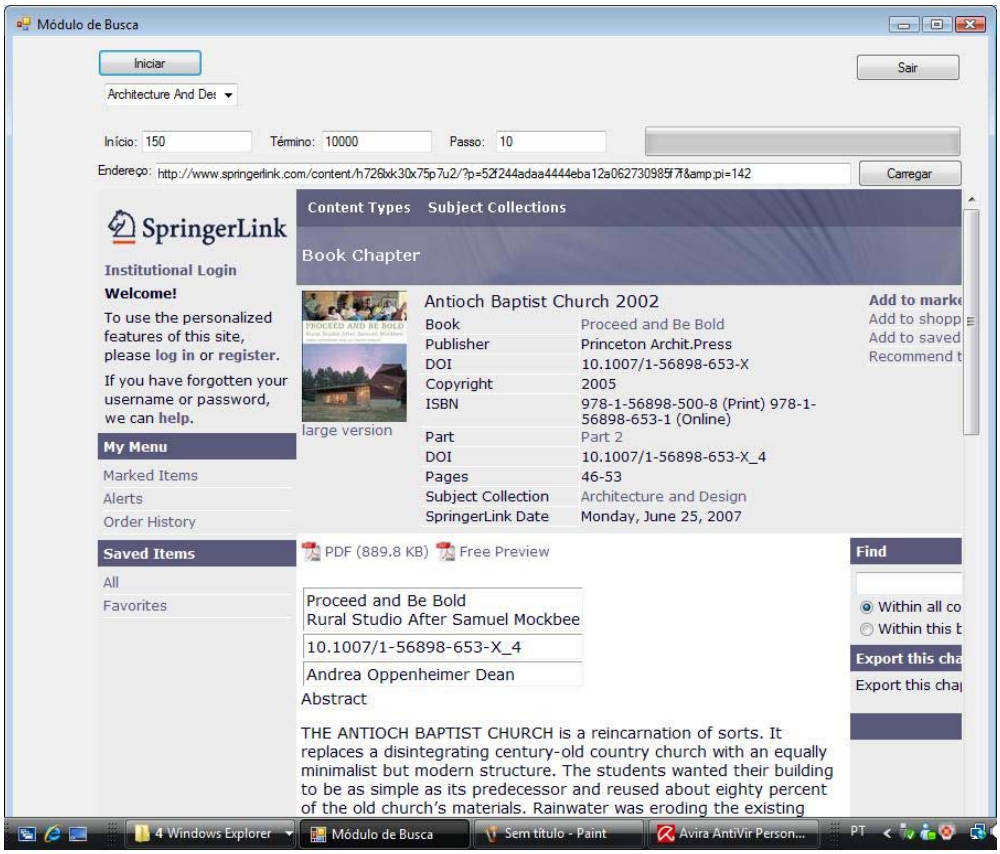


Figura 3.7 – Módulo em modo de Aquisição e Armazenamento.

3.3. O Módulo Preparador

O Módulo Preparador é uma ferramenta de mineração de textos, capaz de carregar os arquivos armazenados pelo módulo de Busca. Os arquivos armazenados podem ser transportados para outras máquinas diferentes, com o intuito de minimizar o tempo gasto para esta tarefa. As máquinas que farão a preparação de dados podem ser as mesmas que fizeram a busca, mas se usadas em conjunto, um grupo pode fazer a busca e aquisição enquanto outro pode realizar a preparação. Uma vez que os arquivos estão em um diretório local, a tarefa se realiza de forma mais rápida, sem a necessidade de uso de banda da rede local. Os arquivos podem ser acessados via rede, mas o acesso a cada um deles acarretará uso de banda da rede e prejudicará o desempenho durante o processo de inclusão dos registros no servidor de banco de dados. As tarefas deste módulo são extrair os dados referentes a cada publicação, preparar os dados, montar as estruturas dos vetores de palavras e ocorrências do resumo (*Abstract*) e armazenar os registros de forma estruturada no gerenciador de banco de dados para formar a base de

dados a ser utilizada pelo módulo Pesquisador. As principais funções do módulo Preparador são:

- Carregar os arquivos armazenados no diretório local;
- Verificar os arquivos para evitar duplicidade na base de dados;
- Extração dos dados que configuram a publicação, tais como título, data de publicação, etc.;
- Armazenar os dados da publicação de forma estruturada na base de dados na tabela de artigos;
- Aplicação dos algoritmos de mineração de textos, tais como *Case Folding*, *Stemming* e *Stopwords* sobre o resumo da publicação para obter as palavras relevantes e suas ocorrências;
- Armazenar as palavras relevantes obtidas no resumo de cada publicação respectivamente e as ocorrências de cada uma destas palavras na tabela de palavras que é relacionada com a tabela de artigos na base de dados.

A metodologia apresentada visa aumentar a velocidade de processamento dos diferentes módulos, uma vez que o módulo Preparador mantém em uma base as palavras e suas ocorrências, o módulo Pesquisador precisa apenas carregar os dados da base, sem a necessidade de qualquer preparação. Todos os documentos recebidos serão submetidos ao módulo Preparador, que se divide em seis diferentes tarefas e seu algoritmo é apresentado na figura 3.8:

```
Para cada publicação definidas por assunto escolhido e início e fim
        Abrir o arquivo
        Verificar arquivo
        Extrair os dados para o registro da publicação
        Extrair as palavras relevantes de cada resumo e quantidade de ocorrências
        Armazenar o registro da publicação na tabela de publicações
        Para cada palavra relevante
            Armazenar a palavra e o número de ocorrências na tabela de palavras
        Fim-Para
Fim-Para
```

Figura 3.8 – Algoritmo para Preparação de dados.

As tarefas definidas foram agrupadas e divididas em 3 diferentes partes:

1. Carga e extração de dados:

Esta parte carrega o texto armazenado, verifica o arquivo e extraí os dados. Cada arquivo é então submetido ao de algoritmo de extração de dados, capaz de efetuar a extração dos dados referentes a publicação. Os dados extraídos são armazenados na base de dados, com informações importantes sobre a publicação, tais como:

- Título;
- Série da Publicação;
- Origem da Publicação;
- ISSN Impresso;
- ISSN OnLine;
- Volume;
- Livro;
- Ano dos Direitos Autorais (*copyright*);
- ISBN;
- Páginas;
- Assunto;
- Data da Publicação;
- DOI;
- Resumo (*Abstract*);
- Tipo de publicação;
- Data da aquisição pelo Sistema.

2. Preparação dos Dados:

Esta parte realiza as tarefas de preparação de dados dos resumos das publicações, definidas pelo processo de *Text Mining* para limpeza de palavras irrelevantes, preparação das palavras e extração das palavras relevantes e suas ocorrências no resumo. Estes processos são definidos graficamente através da figura 3.9 e compõem-se de:

- *Case Folding – Todas as palavras são transformadas para sua forma em letras minúsculas, impedindo que a mesma palavra seja representada de forma diferente para as comparações.*

- *Stopwords* - o texto resultante é submetido ao algoritmo capaz de extrair as palavras, eliminando todo tipo de dados que não formem palavras, além de eliminar todas as palavras que estiverem na lista de stopwords (*stoplist*), palavras essas que são irrelevantes para o contexto. A stoplist utilizada é apresentada no apêndice I e contém uma lista destinada ao idioma inglês.

- *Stemming* - É tarefa final da fase de preparação dos dados e consiste na tarefa de *stemming* do documento. Através do uso do algoritmo de *stemming* de Porter[22], onde cada palavra é reduzida ao seu radical.

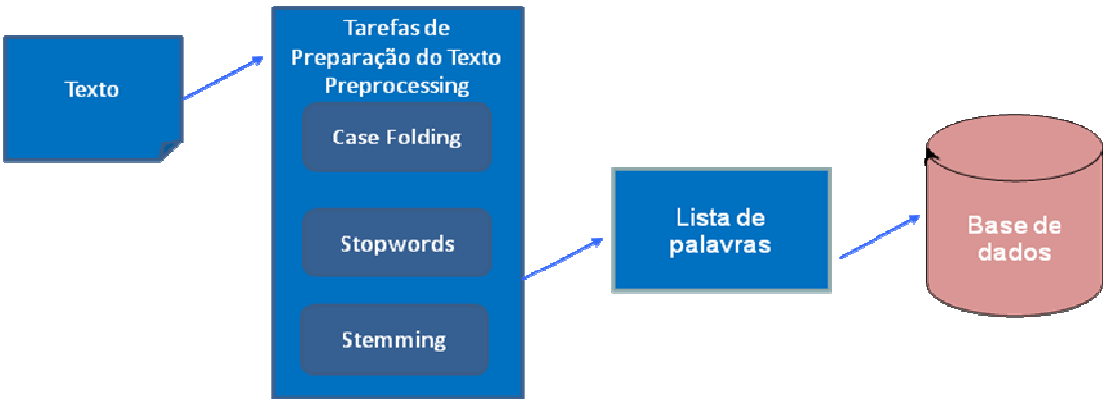


Figura 3.9 – Processo de preparação de dados usado em *Text Mining*.

O processo é definido pelo algoritmo apresentado na figura 3.10 e além das tarefas de mineração de textos apresentadas anteriormente, ainda é necessária a retirada da pontuação das frases.

```
Para cada palavra do resumo
    Eliminar pontuações
    Se palavra possuir mais de uma letra
        Eliminar os espaços anteriores e posteriores
        Transformar todas as letras para minúsculas
        Aplicar o algoritmo de StopWord
        Aplicar o algoritmo de Stemming
    Fim-se
Fim-para
```

Figura 3.10 – Algoritmo para obtenção das palavras relevantes do resumo.

3. Armazenamento de Dados:

Após esta preparação preliminar realizada pela Preparação dos Dados, o vetor resultante contendo apenas as palavras relevantes é compactado, eliminando-se as duplicidades de palavras e calculando o total de ocorrências de cada palavra. A figura 3.11 apresenta o algoritmo para a compactação do vetor de palavras relevantes e suas respectivas ocorrências.

```
Ordenar o vetor de palavras
Para cada palavra atual do vetor
    Se a palavra atual for igual a próxima
      Eliminar a próxima palavra do vetor
      Incrementar o número de ocorrências da palavra atual em 1
    Fim-se
    Ler próxima palavra do vetor
Fim-para
```

Figura 3.11 – Algoritmo para compactação do vetor de palavras relevantes do resumo.

Ao término da compactação do vetor, é criado um registro na tabela de artigos com os dados obtidos na tarefa 1 e cada palavra deve ser armazenada na tabela de palavras a partir do vetor compactado. Cada palavra deve ser associada ao seu número de ocorrências no texto e identificada pelo código de identificação do registro do artigo ao qual pertence.

3.4. O Módulo Pesquisador

O Módulo Pesquisador é responsável pelo processamento final e calcula as distâncias entre o resumo definido pelo usuário e as publicações armazenadas no servidor. Este módulo também é responsável pelo cálculo do Índice de Aproximação Estatístico (IAE), que é calculado em função das ocorrências das palavras definidas pelo assunto e o ajuste dos pesos. O perfil então é definido pelo resumo e a definição das palavras relevantes ao assunto e seus respectivos pesos, que são definidos pelo usuário. O resumo sofre os mesmos processos realizados sobre os resumos das publicações são submetidos e já descritos nas partes 2 e 3 da seção 3.3 e gera um vetor de formato semelhante aos das publicações. O perfil é definido durante o cadastramento do usuário no sistema.

As distâncias são calculadas entre o vetor referente a cada publicação armazenada na base de dados e o vetor produzido pelo resumo definido pelo usuário. O IAE também é calculado neste processo e os resultados são armazenados na tabela de distâncias que representa a distância e o IAE do perfil e cada publicação na base de dados. Uma lista de n palavras ordenada pelas menores distâncias calculadas é gerada, identificando as publicações de maior relevância para o usuário e o valor do IAE no qual é usado para reclassificar a lista em nova ordem. A lista poderá ser parametrizada de forma a atender melhor as necessidades do usuário, bastando informar, por exemplo, a quantidade de itens a serem apresentados na lista.

3.5. Modelagem da Ferramenta Desenvolvida

O diagrama de pacotes e seus respectivos diagramas de casos de uso a seguir, nas figuras 3.12, 3.13, 3.14 e 3.15, descreve o cenário que apresenta as funcionalidades do sistema do ponto de vista do usuário. A modelagem foi feita a partir dos modelos propostos pela UML, linguagem de modelagem unificada (*Unified Modeling Language*).

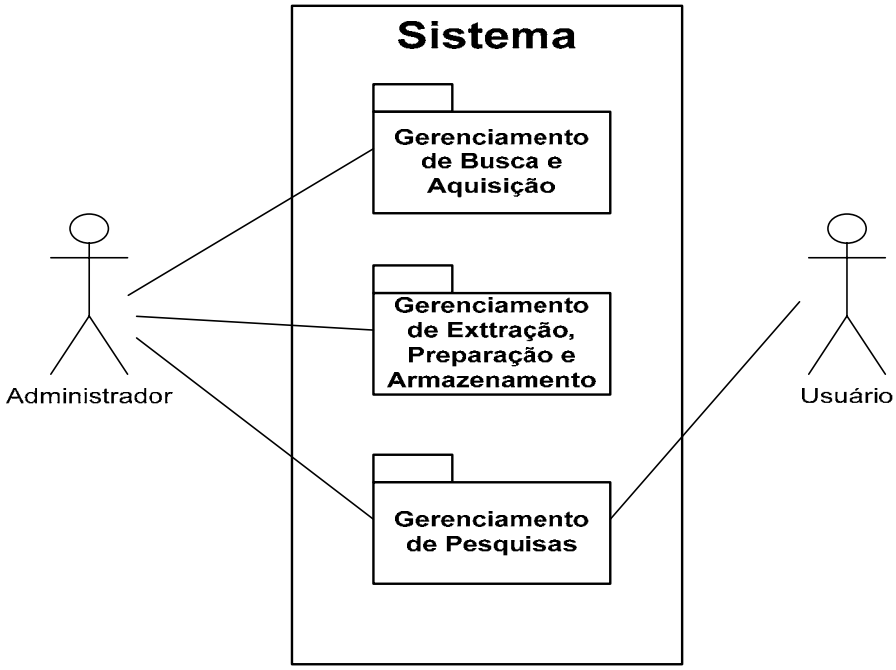


Figura 3.12 – Diagrama de Pacotes.

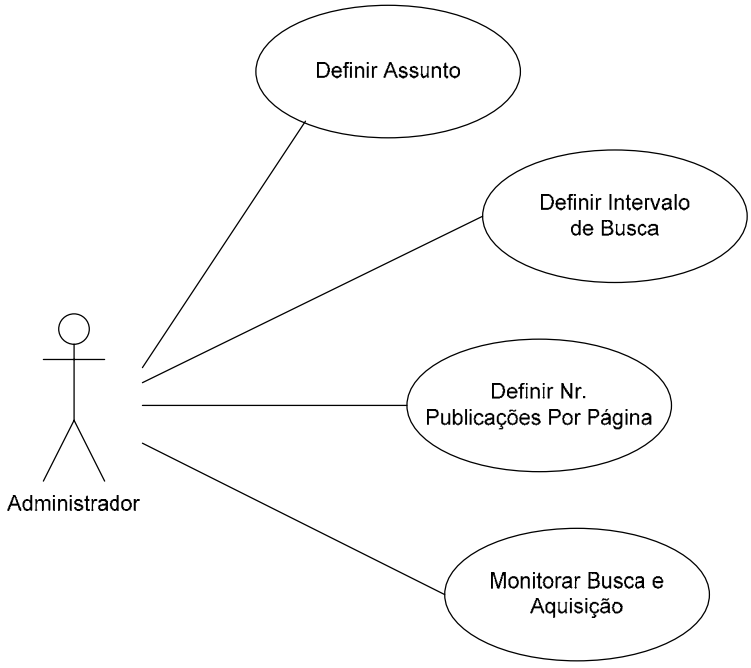


Figura 3.13 – Caso de Uso: Gerenciamento de Busca e Aquisição

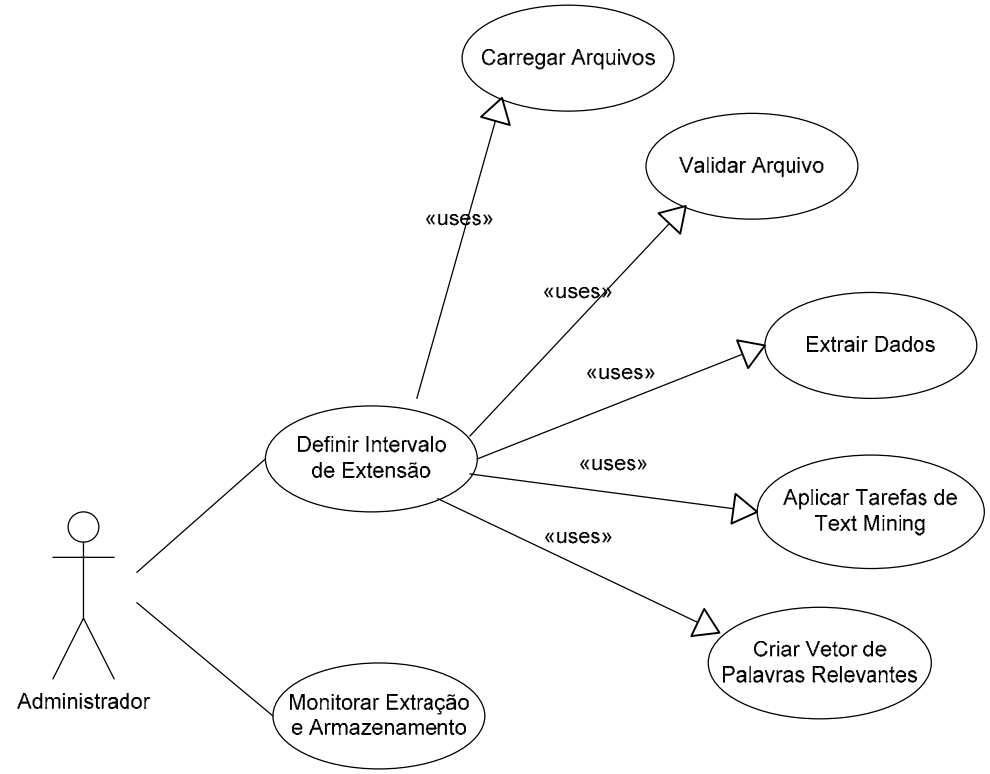


Figura 3.14 – Caso de Uso: Gerenciamento Extração, Preparação e Armazenamento

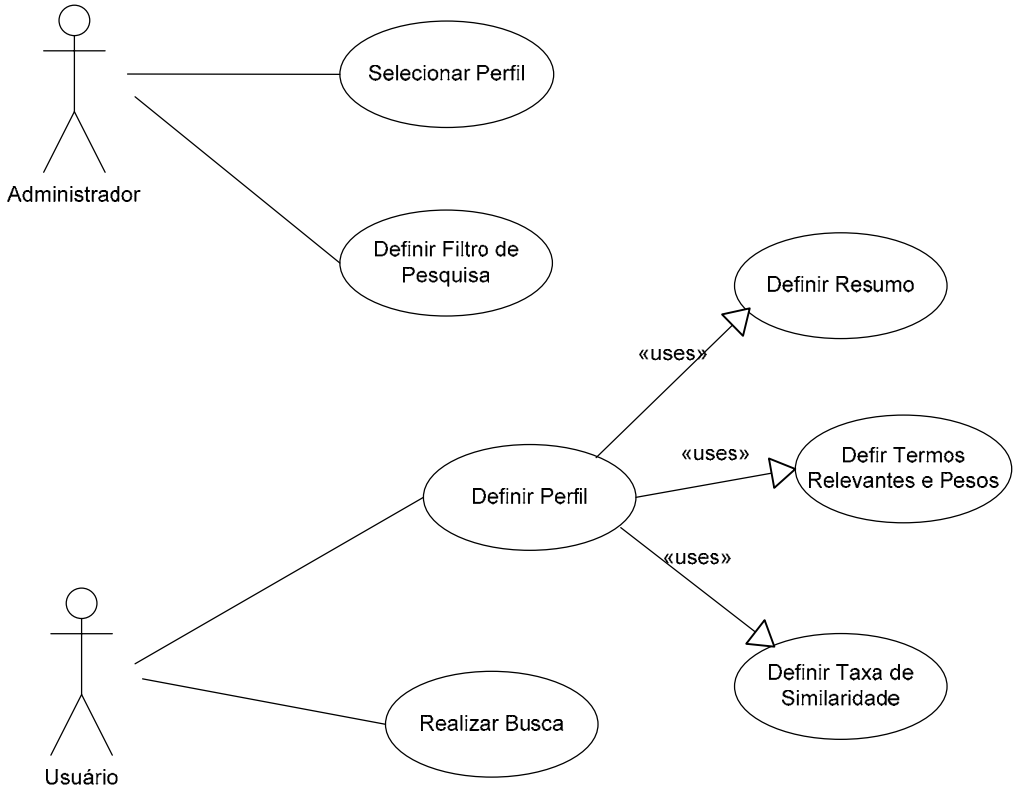


Figura 3.15 – Caso de Uso: Gerenciamento de Pesquisas

3.6. Arquitetura da Ferramenta Desenvolvida

A metodologia visa o emprego das diferentes tarefas definidas até aqui em dois momentos diferentes. No primeiro momento as tarefas de Busca e Preparação são executadas em diferentes máquinas para aumentar o desempenho do sistema. Como cada uma destas tarefas foi projetada para trabalhar de forma sequencial, elas podem ser utilizadas de forma simultânea. Em máquinas com um único processador, os testes comprovaram que seu desempenho só começa a degradar com o uso de mais de quatro instâncias simultâneas, executando de forma concorrente. Isto é comprovado pelos gráficos de desempenho de processamento para os módulos: Buscador, Preparador e Pesquisador que podem ser observados nas figuras apresentadas nas próximas seções. Já em máquinas com mais de um núcleo, a metodologia permite que estas instâncias sejam executadas de formas paralelas e concorrentes, o que permite que até quatro instâncias sejam executadas por núcleo sem que ocorra degradação do sistema. Existe ainda a possibilidade de se executar estas tarefas em um conjunto de máquinas com mais de um núcleo, o que permite um alto desempenho para o sistema. A figura 3.16 apresenta uma configuração de máquinas para a realização das tarefas de Busca, Preparação e Pesquisa, sendo que várias máquinas auxiliares podem ser usadas. A metodologia teve seu desenvolvimento implementado através de um conjunto de ferramentas criadas em linguagem C# no ambiente *Visual Studio* 2008.

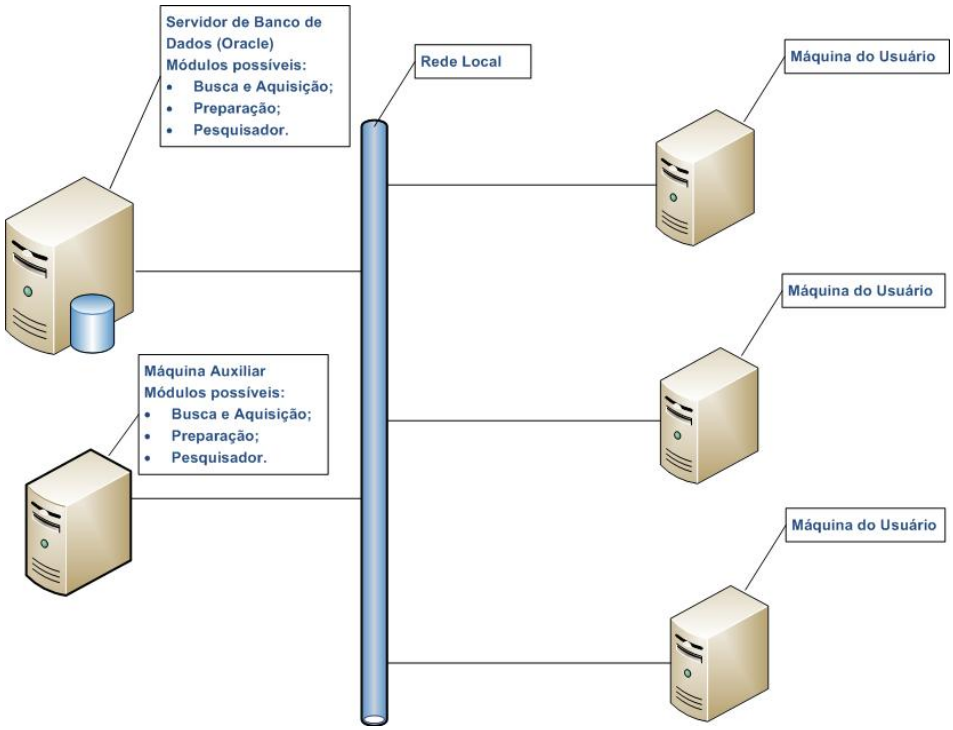


Figura 3.16 – Arquitetura para Busca, Preparação e Pesquisa.

3.6.1 Arquitetura do Módulo Buscador

O módulo Buscador é usado inicialmente para adquirir um grande conjunto de dados e posteriormente para adquirir apenas novas publicações disponíveis. A aquisição destas novas publicações não interfere nos demais módulos, uma vez que esta tarefa pode ser feita em qualquer outra máquina, não sendo necessária a utilização do servidor para tal. Isso permite o aumento do desempenho da metodologia, pois não há impacto algum sobre as demais tarefas da metodologia. A aquisição pode inclusive ser feita em outro local geográfico, bastando que ao final os arquivos adquiridos sejam transportados para a rede local ao servidor. Isto proporciona que a aquisição possa ser feita em local de alta capacidade de download e distribuído por um conjunto de máquinas, já que o acesso sequencial permite que a aquisição seja feita em paralelo por estas máquinas. Na fase inicial, a quantidade de publicações a ser adquirida é muito grande e o principal objetivo desta arquitetura é o desempenho dos processos.

A máquina para a execução deste módulo não requer sistema operacional sofisticado, nem qualquer tipo de dispositivo especial, é apenas necessário que a máquina esteja conectada a internet e seu sistema operacional tenha suporte a aplicações desenvolvidas em *Visual Studio* da *Microsoft*. Ela possui um objeto de navegação (*WebBrowser*) que obtém os links e os arquivos das publicações. Deve-se escolher o assunto e definir de forma seqüencial o número da publicação inicial e o número da publicação final. Esta forma possibilita que outras instâncias da aplicação executem simultaneamente em uma mesma máquina ou em máquinas diferentes, aumentando a velocidade de aquisição de dados. Um conjunto de máquinas com acesso a internet pode realizar a tarefa de aquisição de dados em um tempo muito menor e como a tarefa é automatizada, basta apenas definir o assunto que cada uma vai buscar e definir o número da publicação inicial e final. Enquanto uma instancia obtém publicações do número 1 até 10.000, outra pode iniciar de 10.001 até 20.000 e assim por diante. Nos testes por vezes foram usadas cinco máquinas com pelo menos 3 instâncias executando simultaneamente para adquirir os dados. A média foi de mil publicações por hora por instância. A figura 3.17 apresenta o gráfico de desempenho para a aquisição de dados. Pode-se prever que 5 máquinas com 3 instâncias cada pode adquirir 500.000 publicações, conforme apresentado na fórmula 3.1, assim permite-se a obtenção dos dados definidos para o estudo de caso em aproximadamente 34 horas. O uso de mais

máquinas permite que o tempo de aquisição seja menor e o uso de um laboratório com 20 máquinas durante um dia pode adquirir os dados citados em pouco mais de 8 horas.

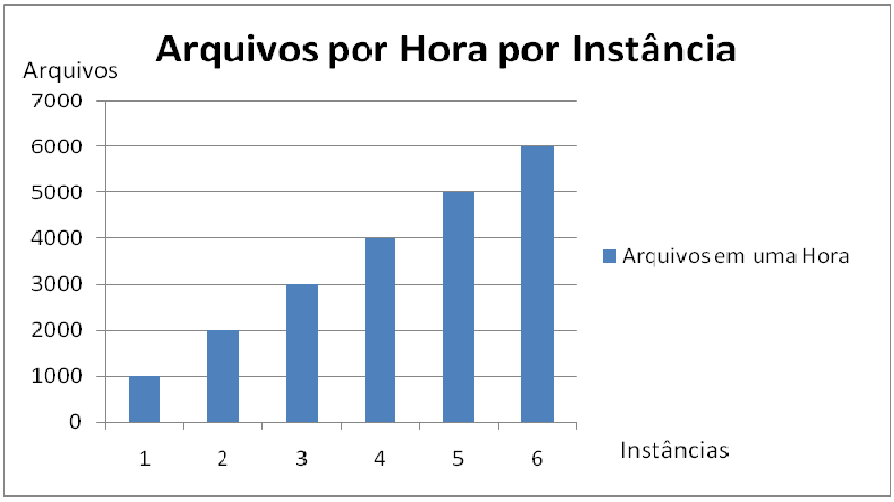


Figura 3.17 – Gráfico de desempenho para aquisição de dados – Módulo Buscador.

Publicações = 500.000

Máquinas = 5

Instâncias = 3

Processamento por hora (proc) = 1.000

$$\text{Tempo Total (h)} = \frac{\text{publicações}}{\text{Máquinas} * \text{Instância} * \text{Processamento}} \rightarrow \frac{500.000}{5 * 3 * 1.000} = \sim 34 \text{ hs.}$$

Fórmula 3.1 – Cálculo do tempo para a aquisição dos dados definidos.

A tarefa de busca e aquisição de dados pode ser feita independentemente das demais tarefas. O projeto visa á inclusão de novas publicações em qualquer momento, sendo necessária apenas a atualização da pesquisa do usuário sobre os novos dados e não necessitando que todo o processo seja repetido. Os módulos tem independência de processamento e de máquina, uma mesma pesquisa não precisa ser refeita, pois a lista de publicações relevantes uma vez pronta pode ser atualizada apenas em função das novas distâncias calculadas sobre as novas publicações incluídas na base de dados. É apenas necessário calcular as distâncias entre o perfil do usuário e as novas publicações e re-gerar a lista de publicações relevantes levando em conta os menores valores das distâncias. Isso pode ser obtido por uma simples consulta à base de dados, ordenada pelas menores distâncias.

3.6.2 Arquitetura do Módulo Preparador

O módulo Preparador deve ser executado em rede local por uma máquina auxiliar com acesso ao servidor da base de dados, ou no próprio servidor. A sua forma sequencial de acesso aos arquivos das publicações possibilita que várias máquinas ou instâncias diferentes sejam executadas simultaneamente, da mesma forma que o módulo Buscador, aumentando o desempenho da metodologia proposta. Isto permite que enquanto outras máquinas façam a tarefa de aquisição, outras máquinas posicionadas em rede com o servidor possam alimentar a base de dados, sem nenhum tipo de interferência ou concorrência entre elas. Na fase inicial, a quantidade de publicações a ser preparada é muito grande e objetivo desta arquitetura é o de aumentar o desempenho do sistema. O gargalo possível está ligado ao trafego de dados na rede e a capacidade de atendimento do servidor. A figura 3.14 apresenta o desempenho desta tarefa e como ela tem desempenho 10 (dez) vezes superior a tarefa de aquisição, é aconselhável ter disponível 10% (dez por cento) das máquinas utilizadas para a aquisição.

A máquina para a execução deste módulo também não requer sistema operacional sofisticado, nem qualquer tipo de dispositivo especial, é apenas necessário que a máquina esteja conectada a rede local e com acesso ao servidor de banco de dados, além de ter suporte a aplicações desenvolvidas em *Visual Studio* da *Microsoft.* Ela possui uma tela para monitoramento e acompanhamento do processo. Deve-se apenas definir de forma seqüencial o número da publicação inicial e o número da publicação final para as publicações que serão processadas. O processo é automatizado e o módulo possibilita que outras instâncias executem simultaneamente em uma mesma máquina ou em máquinas diferentes, aumentando a velocidade de preparação de dados. Um conjunto de máquinas com acesso a rede local pode realizar a tarefa de preparação de dados em um tempo muito menor e como a tarefa é automatizada, basta apenas definir o número da publicação inicial e final. Enquanto uma instância prepara publicações do número 1 até 30.000, outra pode iniciar de 30.001 até 60.000 e assim pode diante. Nos testes por vezes foram usadas duas máquinas executando 2 instâncias simultaneamente na preparação dos dados. A média foi de dez mil publicações por hora por instância. A figura 3.18 apresenta o gráfico de desempenho para a preparação de dados. Pode-se prever que 2 máquinas com 2 instâncias cada pode adquirir 500.000

publicações, conforme apresentado na fórmula 3.2, e assim permite-se a preparação dos dados iniciais definidos para o estudo de caso em aproximadamente 125 horas. O uso de mais máquinas permite que o tempo de aquisição seja menor e o uso de um laboratório com 20 máquinas durante um dia com dois núcleos de processamento cada, pode preparar os dados citados em pouco mais de 12,5 horas.

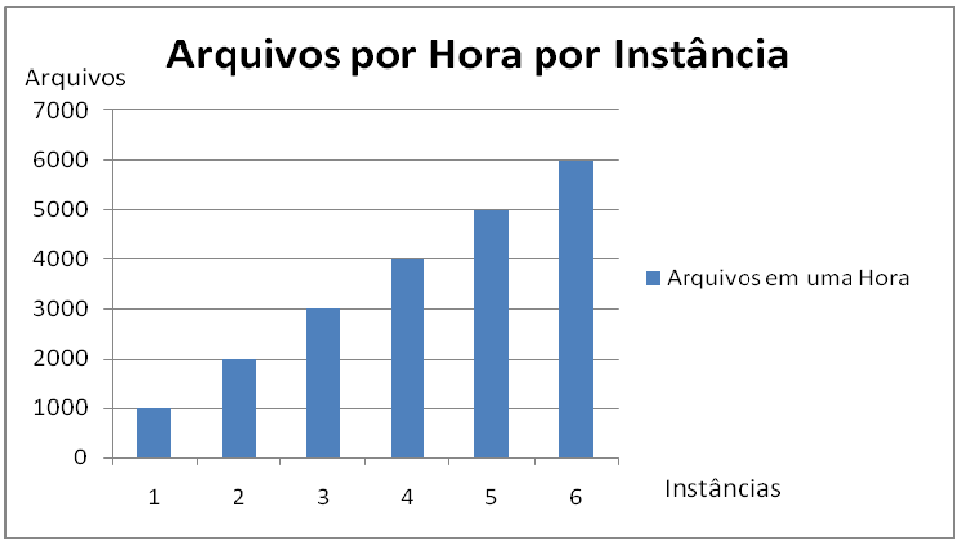


Figura 3.18 – Gráfico de desempenho do módulo de Preparação de dados.

Publicações = ~500.000

Máquinas = 2

Instâncias = 2

Processamento por hora (proc) = 1.000

$$\text{Tempo Total (h)} = \frac{\text{publicações}}{\text{Máquinas} * \text{Instância} * \text{Processamento}} \rightarrow \frac{500.000}{2 * 2 * 1000} = \sim 125 \text{ hs.}$$

Fórmula 3.2 – Cálculo do tempo para a preparação dos dados definidos.

3.6.3 Arquitetura do Módulo Pesquisador

O módulo Pesquisador, assim como o módulo Preparador, deve ser executado em rede local por uma máquina auxiliar com acesso ao servidor da base de dados, ou no próprio servidor. A sua forma sequencial de acesso aos arquivos das publicações possibilita que várias máquinas ou instâncias diferentes sejam executadas simultaneamente, da mesma forma que os módulos: Buscador e Preparador, aumentando o desempenho do sistema proposto. Isto permite que enquanto outras máquinas façam as tarefas de aquisição e preparação, outras máquinas posicionadas em rede com o servidor possam pesquisar a base de dados, tendo como interferência, apenas as máquinas da tarefa de preparação, que concorrerão entre si. Nesta fase, a quantidade de dados armazenados é grande e o objetivo desta arquitetura é aumentar o desempenho do sistema. Como a concorrência através do uso da rede e do servidor é grande, deve-se utilizar uma rede com capacidade de tráfego de pelo menos 1 (um) Gbps e um servidor de alta capacidade de processamento para o atendimento das requisições feitas ao gerenciador de banco de dados. A figura 3.19 apresenta o desempenho desta tarefa e como ela tem desempenho semelhante a da tarefa de aquisição, é aconselhável que seja usada uma máquina para cada usuário com 4 (quatro) instâncias, dividindo a tarefa de acesso a todas as publicações por 4 (quatro) para aumentar o desempenho do sistema.

A máquina para a execução deste módulo requer as mesmas necessidades do módulo anterior. Ela possui uma tela para monitoramento para acompanhamento do processo. Deve-se definir de forma seqüencial o número da publicação inicial e o número da publicação final para as publicações que serão processadas e escolher o usuário para obtenção do perfil. O processo é automatizado e o módulo possibilita que outras instâncias da ferramenta executem simultaneamente em uma mesma máquina ou em máquinas diferentes, aumentando a velocidade da pesquisa. Um conjunto de máquinas com acesso a rede local pode realizar a tarefa de pesquisa nos artigos em um tempo muito menor. Enquanto uma instância obtém publicações de um usuário em todas as publicações, outra pode processar o perfil de outro usuário. Nos testes por vezes foram usadas duas máquinas executando 2 instâncias simultaneamente na pesquisa. A média foi de mil publicações por hora por instância. A figura 3.19 apresenta o gráfico de desempenho para a pesquisa. Pode-se prever que 2 máquinas com 2 instâncias cada pode pesquisar 500.000 publicações, conforme apresentado na fórmula 3.3, e assim permite-se a pesquisa leva aproximadamente 2,5 horas.

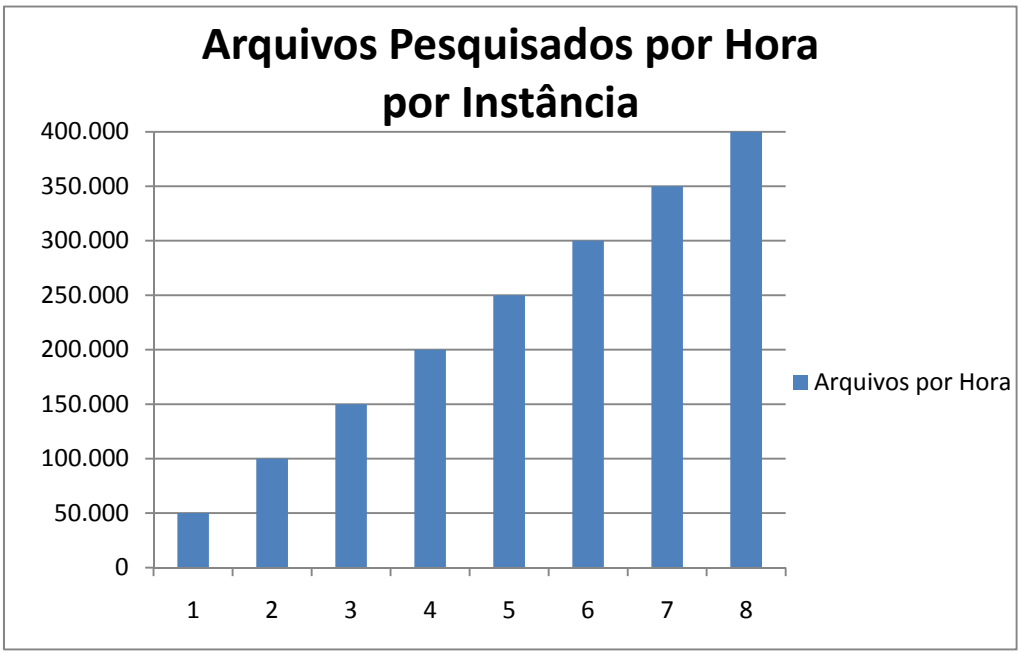


Figura 3.19 – Gráfico de desempenho do Módulo Pesquisador.

Publicações = ~500.000

Máquinas = 2

Instâncias = 2

Processamento por hora (proc) = 50.000

$$\text{Tempo Total (h)} = \frac{\text{publicações}}{\text{Máquinas} * \text{Instância} * \text{Processamento}} \rightarrow \frac{500.000}{2 * 2 * 50.000} = \sim 2{,}5 \text{ hs.}$$

Fórmula 3.3 – Cálculo do tempo para a pesquisa pelo perfil de um usuário.

3.7 Manutenção do Sistema

Após a carga inicial dos dados e das pesquisas iniciais feitas pelos usuários, o sistema entra em uma nova fase, a de manutenção dos dados com a entrada de novas publicações e de atualização das listas de publicações relevantes dos usuários. Estas tarefas podem ser executadas no próprio servidor, uma vez que o número de publicações novas é muito inferior ao número de publicações obtidas na fase inicial do sistema. Para evitar a degradação do sistema que pode ocorrer no caso um novo usuário ser incluído, a tarefa de pesquisa seja feita em uma máquina auxiliar e posteriormente ser atualizada diretamente pelo servidor.

Nesta fase, é mantida no servidor de banco de dados as tarefas de aquisição de novos arquivos, preparação dos dados recém adquiridos e de atualização das pesquisas dos usuários para os cálculos das distâncias para as novas publicações. Apenas para o caso de novos usuários, é aconselhável utilizar uma máquina auxiliar para efetuar a pesquisa inicial do mesmo.

Capítulo 4
Estudo de Caso

O portal da CAPES, é um local de referência para a busca de publicações relevantes. Através dele tem-se acesso a um grande número de sites que disponibilizam milhões de trabalhos científicos. Dentre as várias opções disponíveis, o *Springerlink* (www.springerlink.com) é um dos que se destaca em diversas áreas do conhecimento. O estudo de caso será feito sobre o *Springerlink* e serão utilizados 7 perfis diferentes para avaliação dos resultados. Os resultados são apresentados nas próximas seções deste capítulo separados por perfil. Na figura 4.1 pode-se verificar que são disponibilizadas mais de quatro milhões de publicações classificadas em treze diferentes assuntos. Para os estudos e métricas deste trabalho definiu-se o *Springerlink* como base de publicações e sete diferentes assuntos para a busca e a aquisição de dados, sendo eles definidos na tabela 4.1.

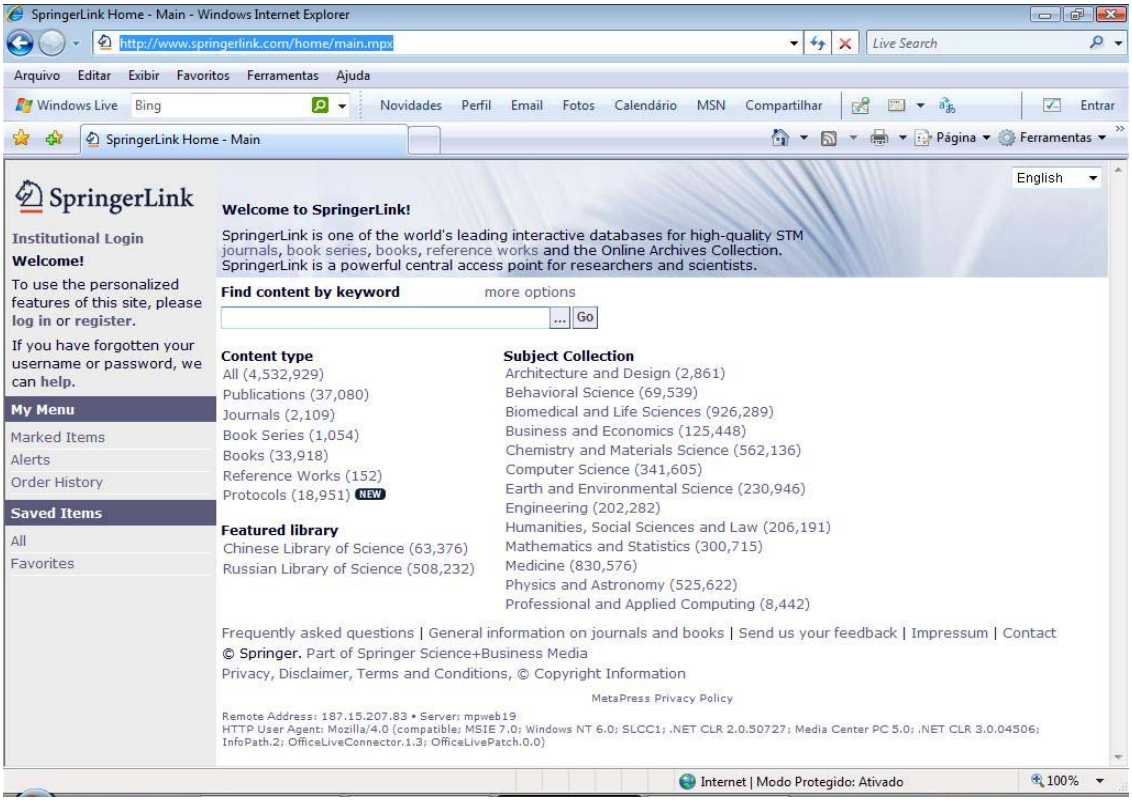


Figura 4.1 – Página principal do site Springerlink: www.springerlink.com

Tabela 4.1 – Assuntos definidos para testes do trabalho de pesquisa.

Assunto	Quantidade de Publicações	Quantidade Definida para os Testes
Architeture and Design	2.861	2.861
Behavioral Science	69.539	20.000
Biomedical and Life Sciences	926.289	20.000
Bussiness and Economics	125.448	20.000
Computer science	341.605	20.000
Medicine	830.576	20.000
Professional and Applied Computing	8.442	8.442
Totais:	**2.304.760**	**111.303**

Fonte: *Springerlink* (setembro/2009).

A busca foi feita sobre estes sete diferentes assuntos, totalizando a aquisição de 111.303 diferentes publicações. A figura 4.1 apresenta um momento da aquisição através do módulo Buscador.

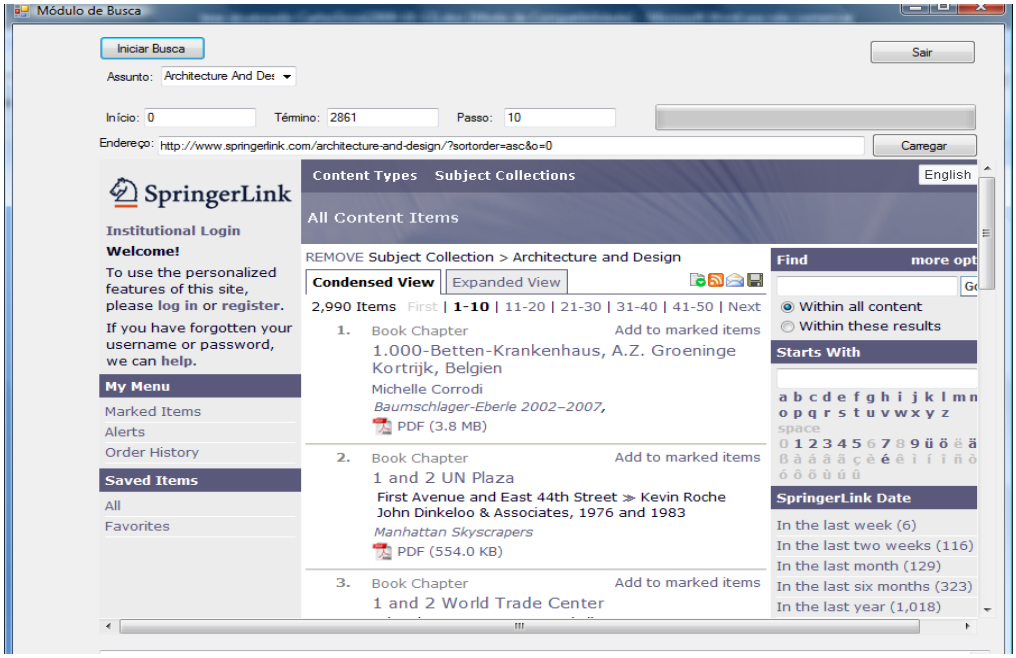


Figura 4.2 – Ferramenta de busca e aquisição no modo busca de links.

A seguir foi realizada a tarefa de preparação dos dados sobre os arquivos adquiridos e preparou-se a base de dados com as informações dos artigos e a relação de palavras relevantes a cada um e suas respectivas ocorrências. A figura 4.3 apresenta o Módulo Preparador realizando a tarefa de preparação de dados.

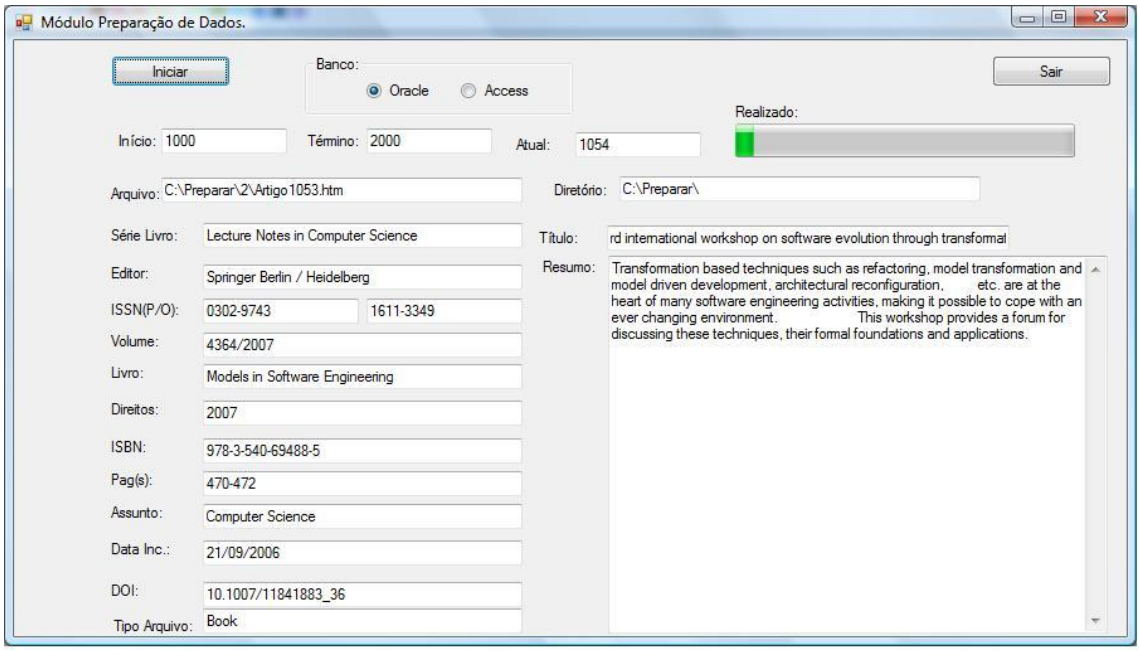


Figura 4.3 – Tela de monitoramento do Módulo Preparador.

Finalmente iniciaram-se os testes sobre a base de dados preparada. A figura 4.4 apresenta o Módulo Pesquisador realizando a tarefa de geração da lista de publicações relevantes.

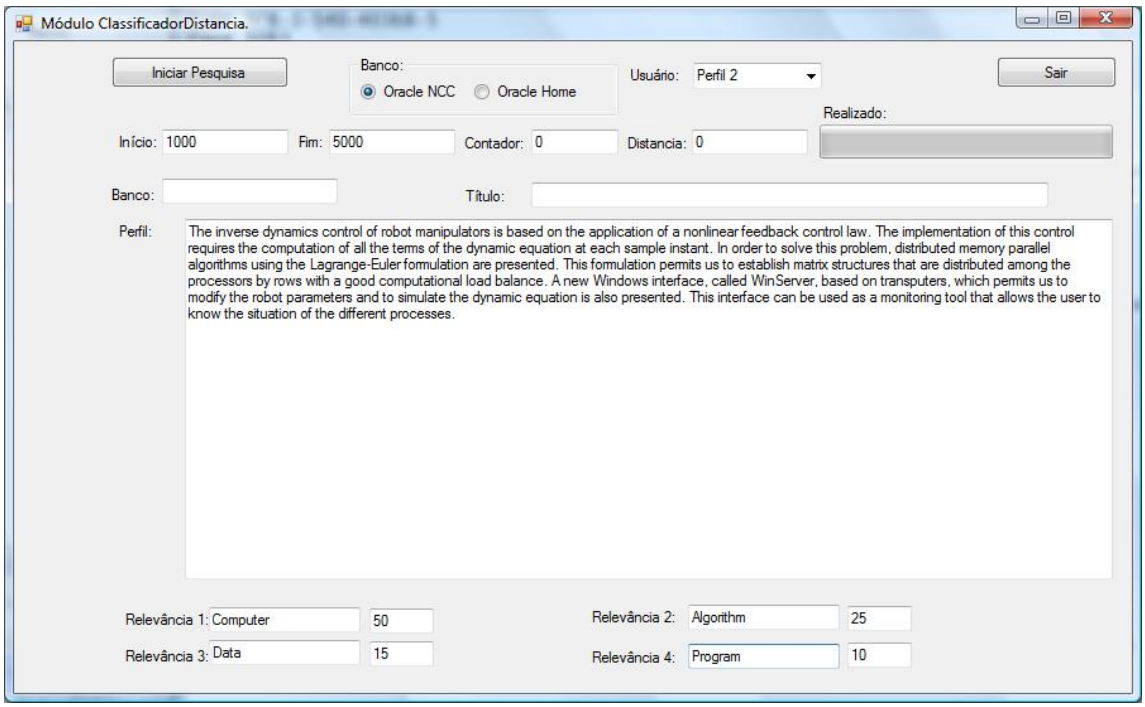


Figura 4.4 – Tela de monitoramento do Módulo Pesquisador.

Os testes foram conduzidos inicialmente com um grupo de oito perfis diferentes, definidos e submetidos ao módulo Pesquisador e que serão apresentados ainda neste capítulo. O primeiro foi definido a partir de uma publicação existente na base para aferir os cálculos de distância, onde o resultado deve apresentar um valor 0 (zero). Os demais perfis foram definidos por publicações dos diferentes assuntos obtidos no mesmo local dos demais, mas não existentes na base, as palavras relevantes foram identificadas através de observações sobre os títulos das publicações referentes a cada assunto. Foram mantidas as mesmas taxas relevância em todos os testes para que as comparações sobre diferentes assuntos pudessem ser feitas. Os resultados obtidos foram analisados através da medição de Precisão e Recuperação para cada um deles, utilizando a classificação através de *KNN* para determinar as distâncias e o Índice de Aproximação Estatística (IAE). Os resultados foram então comparados e as conclusões são apresentadas no capítulo 5. As próximas seções apresentarão os resultados de desempenho da metodologia para os oito perfis definidos. Como as publicações já se encontram previamente classificadas por assunto, os testes levaram em consideração apenas os resumos das publicações e a classe foi utilizada apenas para as avaliações de desempenho. Cada publicação da lista teve seu respectivo assunto obtido durante o processo de extração de dados, sendo assim, as medições de dos índices de Precisão e Recuperação não necessitaram de um especialista.

Nas primeiras observações sobre os resultados, foi detectado que algumas publicações não possuíam resumos expressivos, sejam tanto por inexistência, como por um conjunto muito pequeno de palavras relevantes, o que permitia que estas publicações possuíssem pequena distância, mas nenhuma relevância. Segundo Ebecken ET AL [39], a etapa de preparação dos dados "envolve a seleção dos dados que constituirão a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo dos textos, ou seja, toda a informação que não refletir nenhuma idéia considerada importante poderá ser desprezada." Através de experimentações empíricas ficou definido para este trabalho, que todas as publicações que não possuíssem um mínimo de 25 (vinte e cinco) palavras relevantes fossem descartadas. A tabela 4.2 apresenta o conjunto de publicações obtidas que foram aproveitas neste estudo após a preparação dos dados definidos pela tabela 4.1.

Tabela 4.2 – Publicações aproveitadas no trabalho de pesquisa.

Assunto	Quantidade de Publicações Obtidas	Quantidade Utilizada para Testes
Architeture and Design	2.861	1.672
Behavioral Science	20.000	0
Biomedical and Life Sciences	20.000	12.125
Bussiness and Economics	20.000	7.755
Computer Science	20.000	17.203
Medicine	20.000	2.732
Professional and Applied Computing	8.442	5.259
Totais:	**111.303**	**46.746**

Ao observar a quantidade de publicações utilizadas para análises, pode-se constatar que o assunto "*Behavioral Science*" não apresenta amostras significativas, "*Medicine*" teve um baixo aproveitamento de publicações relevantes com aproveitamento pouco superior a 13%, "*Bussiness and Economics*" teve aproveitamento aproximadamente de 38%. Quanto aos demais assuntos o aproveitamento ficou acima de 50%, destacando o assunto "*Computer Science*", que obteve índice de aproveitamento superior a 86%. Os índices de aproveitamento influenciarão as avaliações dos resultados, uma vez que, por exemplo, perfis montados sobre o assunto "*Behavioral Science*", não apresentarão precisão já que não existem amostras representativas sobre esse assunto.

4.1 Perfil 1 – Aferição da Metodologia Proposta

Para o primeiro perfil, foi definida uma publicação existente na base de dados para aferir a ferramenta desenvolvida, certificando a metodologia proposta. O artigo escolhido foi à primeira publicação adquirida pelo Módulo Preparador, através do assunto: "*Computer Science*". É esperado que ocorra um valor de distância igual a 0 (zero), já que teremos dois resumos iguais. A tabela 4.3 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.4 e 4.5 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística, respectivamente. A relevância é positiva quando o assunto da publicação é "*Computer Science*" e negativo para qualquer outro caso.

Tabela 4.3 – Perfil 1 utilizado na pesquisa.

Assunto: ***Computer Science***	
Título: ***Alpha Scale Spaces on a Bounded Domain***	
Palavras Relevantes:	**Peso**
1ª) *Computer*	50%
2ª) *Algorithm*	25%
3ª) *Data*	15%
4ª) *Program*	10%
Resumo: We consider a scale spaces, a parameterized class (a Î ([0,1]) of scale space representations beyond the well established Gaussian scale space, which are generated by the a th power of the minus Laplace operator on a bounded domain using the Neumann boundary condition. The Neumann boundary condition ensures that there is no grey value flux through the boundary. Thereby no artificial grey values from outside the image affect the evolution proces, which is the case for the a scale spaces on an unbounded domain. Moreover, the connection between the a scale spaces which is not trivial in the unbounded domain case, becomes straightforward The generator of the Gaussian semigroup extends to a compact, self adjoint operator on the Hilbert space \mathbb L 2(W) and therefore it has a complete countable set of eigen functions. Taking the a th power of the Gaussian generator simply boils down to taking the a th power of the corresponding eigenvalues. Consequently, all a scale spaces have exactly the same eigen modes and can be implemented simultaneously as scale dependent Fourier series. The only difference between them is the (relative) contribution of each eigen mode to the evolution proces. By introducing the notion of (non dimensional) relative scale in each a scale space, we are able to compare the various a scale spaces. The case a = 0.5, where the generator equals the square root of the minus Laplace operator leads to Poisson scale space, which is at least as interesting as Gaussian scale space and can be extended to a (Clifford) analytic scale space.	

Tabela 4.4 – Perfil 1, classificação por *KNN*. Assunto: "*Computer Science*".

Doc.	*Distância*	*Assunto*	*Relevância*	*Precisão*	*Recup.*
1	**00,00**	**Computer Science**	**+**	**100%**	**3%**
2	15,59	Business and Economics	-	50%	3%
3	17,66	Computer Science	+	67%	7%
4	17,83	Computer Science	+	75%	10%
5	17,83	Computer Science	+	80%	14%
6	19,67	Computer Science	+	83%	17%
7	19,70	Computer Science	+	86%	21%
8	19,80	Computer Science	+	88%	24%
9	19,82	Computer Science	+	89%	28%
10	19,82	Computer Science	+	90%	31%
11	19,95	Computer Science	+	91%	34%
12	20,00	Architecture and Design	-	83%	34%
13	20,02	Architecture and Design	-	77%	34%
14	20,10	Computer Science	+	79%	38%
15	20,15	Computer Science	+	80%	41%
16	20,25	Computer Science	+	81%	45%
17	20,25	Computer Science	+	82%	48%
18	20,40	Computer Science	+	83%	52%
19	20,40	Biomedical and Life Sciences	-	79%	52%
20	20,42	Computer Science	+	80%	55%
21	20,47	Business and Economics	-	76%	55%
22	20,49	Professional and Applied Computing	-	73%	55%
23	20,54	Computer Science	+	74%	59%
24	21,02	Business and Economics	-	71%	59%
25	21,54	Computer Science	+	72%	62%
26	21,84	Architecture and Design	-	69%	62%
27	21,86	Professional and Applied Computing	-	67%	62%
28	21,91	Computer Science	+	68%	66%
29	22,07	Computer Science	+	69%	69%
30	22,11	Computer Science	+	70%	72%
31	22,14	Computer Science	+	71%	76%
32	22,20	Computer Science	+	72%	79%
33	22,20	Architecture and Design	-	70%	79%
34	22,25	Computer Science	+	71%	83%
35	22,25	Computer Science	+	71%	86%
36	22,29	Architecture and Design	-	69%	86%
37	22,29	Architecture and Design	-	68%	86%
38	22,29	Computer Science	+	68%	90%
39	22,32	Architecture and Design	-	67%	90%
40	22,34	Architecture and Design	-	65%	90%
41	22,36	Architecture and Design	-	63%	90%
42	22,36	Computer Science	+	64%	93%
43	22,38	Architecture and Design	-	63%	93%
44	22,41	Business and Economics	-	61%	93%
45	22,43	Business and Economics	-	60%	93%
46	22,49	Business and Economics	-	59%	93%
47	22,49	Business and Economics	-	57%	93%
48	22,52	Computer Science	+	58%	97%
49	22,54	Computer Science	+	59%	100%
50	22,54	Business and Economics	-	58%	100%

Tabela 4.5 – Perfil 1, reclassificação por IAE. Assunto: "*Computer Science*".

Doc.	*Distância*	*IAE*	*Assunto*	*Relevânica*	*Precisão*	*Recup.*
1	21,91	2,5	Computer Science	+	100%	3%
2	20,15	2	Computer Science	+	100%	7%
3	20,54	1	Computer Science	+	100%	10%
4	19,82	0,75	Computer Science	+	100%	14%
5	22,25	0,75	Computer Science	+	100%	17%
6	19,70	0,5	Computer Science	+	100%	21%
7	19,95	0,5	Computer Science	+	100%	24%
8	20,10	0,5	Computer Science	+	100%	28%
9	20,47	0,5	Business and Economics	-	89%	28%
10	22,20	0,5	Architecture and Design	-	80%	28%
11	22,36	0,5	Computer Science	+	82%	31%
12	22,43	0,5	Business and Economics	-	75%	31%
13	15,59	0,25	Business and Economics	-	69%	31%
14	20,25	0,25	Computer Science	+	71%	34%
15	22,07	0,25	Computer Science	+	73%	38%
16	22,29	0,25	Computer Science	+	75%	41%
17	22,52	0,15	Computer Science	+	76%	45%
18	22,54	0,15	Computer Science	+	78%	48%
19	20,40	0,1	Biomedical and Life Sciences	-	74%	48%
20	20,49	0,1	Professional and Applied Computing	-	70%	48%
21	**00,00**	**0**	**Computer Science**	**+**	**71%**	**52%**
22	17,66	0	Computer Science	+	73%	55%
23	17,83	0	Computer Science	+	74%	59%
24	17,83	0	Computer Science	+	75%	62%
25	19,67	0	Computer Science	+	76%	66%
26	19,80	0	Computer Science	+	77%	69%
27	19,82	0	Computer Science	+	78%	72%
28	20,00	0	Architecture and Design	-	75%	72%
29	20,02	0	Architecture and Design	-	72%	72%
30	20,25	0	Computer Science	+	73%	76%
31	20,40	0	Computer Science	+	74%	79%
32	20,42	0	Computer Science	+	75%	83%
33	21,02	0	Business and Economics	-	73%	83%
34	21,54	0	Computer Science	+	74%	86%
35	21,84	0	Architecture and Design	-	71%	86%
36	21,86	0	Professional and Applied Computing	-	69%	86%
37	22,11	0	Computer Science	+	70%	90%
38	22,14	0	Computer Science	+	71%	93%
39	22,20	0	Computer Science	+	72%	97%
40	22,25	0	Computer Science	+	73%	100%
41	22,29	0	Architecture and Design	-	71%	100%
42	22,29	0	Architecture and Design	-	69%	100%
43	22,32	0	Architecture and Design	-	67%	100%
44	22,34	0	Architecture and Design	-	66%	100%
45	22,36	0	Architecture and Design	-	64%	100%
46	22,38	0	Architecture and Design	-	63%	100%
47	22,41	0	Business and Economics	-	62%	100%
48	22,49	0	Business and Economics	-	60%	100%
49	22,49	0	Business and Economics	-	59%	100%
50	22,54	0	Business and Economics	-	58%	100%

Foram encontradas 29 (vinte e nove) publicações sobre o assunto "*Computer Science*" e o gráfico da figura 4.5 representa a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve um ganho significativo, uma vez que inicialmente se tinha Precisão máxima (100%) apenas no 1º documento e a partir da reclassificação por IAE, este índice se manteve máximo até o 8º documento. A Recuperação que atingia o índice máximo apenas no 49º documento passou a atingir a recuperação máxima já no 40º documento. Foi detectado o documento idêntico com distância 0 (zero), o que comprova a eficácia da metodologia quanto a recuperação de informação.

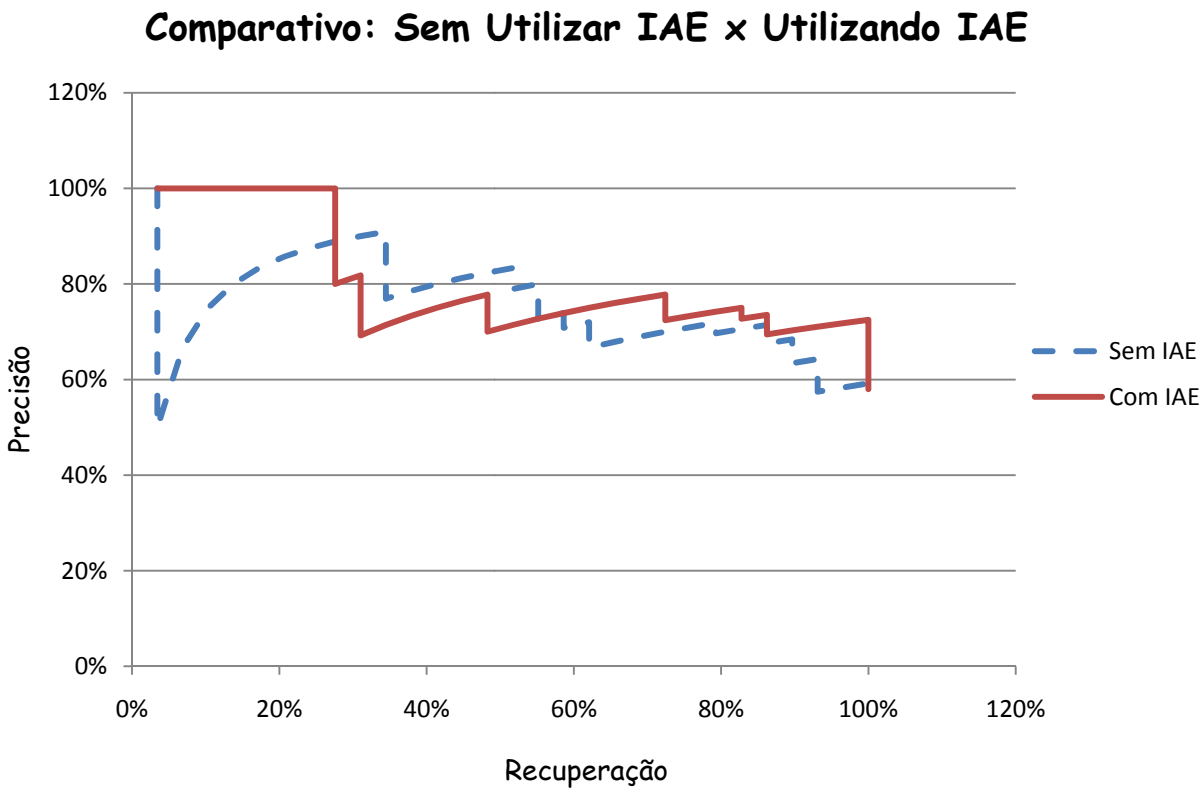


Figura 4.5 – Perfil 1: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma grande melhoria tanto na Precisão quanto na Recuperação de documentos.

4.2 Perfil 2 – Assunto: Computer Science

Para o segundo perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Computer Science*". A tabela 4.6 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.7 e 4.8 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Computer Science*" e negativo para os demais casos nas tabelas 4.7 e 4.8.

Tabela 4.6 – Perfil 2 utilizado na pesquisa. Assunto: "*Computer Science*".

Assunto: ***Computer Science***	
Título: ***An interface based on transputers to simulate the dynamic equation of robot manipulators using parallel computing***	
Palavras Relevantes:	**Peso**
1ª) *Computer*	50%
2ª) *Algorithm*	25%
3ª) *Data*	15%
4ª) *Program*	10%
Resumo: The inverse dynamics control of robot manipulators is based on the application of a nonlinear feedback control law. The implementation of this control requires the computation of all the terms of the dynamic equation at each sample instant. In order to solve this problem, distributed memory parallel algorithms using the Lagrange-Euler formulation are presented. This formulation permits us to establish matrix structures that are distributed among the processors by rows with a good computational load balance. A new Windows interface, called WinServer, based on transputers, which permits us to modify the robot parameters and to simulate the dynamic equation is also presented. This interface can be used as a monitoring tool that allows the user to know the situation of the different processes.	

Tabela 4.7 – Perfil 2, classificação por *KNN*. Assunto: "*Computer Science*".

Doc.	*Distânca*	*Assunto*	*Relevância*	*Precisão*	*Recup.*
1	10,49	Computer Science	+	100%	3%
2	10,58	Computer Science	+	100%	7%
3	10,72	Computer Science	+	100%	10%
4	10,72	Professional and Applied Computing	-	75%	10%
5	10,82	Professional and Applied Computing	-	60%	10%
6	10,91	Computer Science	+	67%	14%
7	10,91	Computer Science	+	71%	17%
8	10,95	Computer Science	+	75%	21%
9	11,00	Computer Science	+	78%	24%
10	11,00	Business and Economics	-	70%	24%
11	11,05	Medicine	-	64%	24%
12	11,05	Computer Science	+	67%	28%
13	11,05	Computer Science	+	69%	31%
14	11,09	Computer Science	+	71%	34%
15	11,14	Business and Economics	-	67%	34%
16	11,14	Computer Science	+	69%	38%
17	11,14	Professional and Applied Computing	-	65%	38%
18	11,14	Computer Science	+	67%	41%
19	11,14	Computer Science	+	68%	45%
20	11,14	Computer Science	+	70%	48%
21	11,14	Computer Science	+	71%	52%
22	11,14	Computer Science	+	73%	55%
23	11,18	Business and Economics	-	70%	55%
24	11,18	Business and Economics	-	67%	55%
25	11,18	Business and Economics	-	64%	55%
26	11,18	Medicine	-	62%	55%
27	11,18	Medicine	-	59%	55%
28	11,18	Medicine	-	57%	55%
29	11,18	Computer Science	+	59%	59%
30	11,18	Computer Science	+	60%	62%
31	11,18	Computer Science	+	61%	66%
32	11,18	Computer Science	+	63%	69%
33	11,18	Computer Science	+	64%	72%
34	11,18	Computer Science	+	65%	76%
35	11,18	Medicine	-	63%	76%
36	11,22	Business and Economics	-	61%	76%
37	11,22	Business and Economics	-	59%	76%
38	11,22	Biomedical and Life Sciences	-	58%	76%
39	11,22	Computer Science	+	59%	79%
40	11,22	Medicine	-	58%	79%
41	11,22	Computer Science	+	59%	83%
42	11,22	Computer Science	+	60%	86%
43	11,22	Computer Science	+	60%	90%
44	11,22	Computer Science	+	61%	93%
45	11,22	Computer Science	+	62%	97%
46	11,22	Medicine	-	61%	97%
47	11,27	Business and Economics	-	60%	97%
48	11,27	Business and Economics	-	58%	97%
49	11,27	Computer Science	+	59%	100%
50	11,27	Medicine	-	58%	100%

Tabela 4.8 – Perfil 2, reclassificação por IAE. Assunto: "*Computer Science*".

Doc.	*Distância*	*IAE*	*Assunto*	*Relevância*	*Precisão*	*Recup.*
1	11,22	1,5	Computer Science	+	100%	3%
2	10,91	1,1	Computer Science	+	100%	7%
3	11,18	1,1	Computer Science	+	100%	10%
4	11,22	1	Computer Science	+	100%	14%
5	11,14	0,85	Computer Science	+	100%	17%
6	10,91	0,75	Computer Science	+	100%	21%
7	11,14	0,75	Computer Science	+	100%	24%
8	11,18	0,75	Computer Science	+	100%	28%
9	11,18	0,75	Computer Science	+	100%	31%
10	11,18	0,75	Computer Science	+	100%	34%
11	11,22	0,65	Computer Science	+	100%	38%
12	11,05	0,6	Computer Science	+	100%	41%
13	10,58	0,5	Computer Science	+	100%	45%
14	11,14	0,5	Computer Science	+	100%	48%
15	11,18	0,5	Computer Science	+	100%	52%
16	11,22	0,5	Computer Science	+	100%	55%
17	11,27	0,5	Business and Economics	-	94%	55%
18	11,14	0,25	Business and Economics	-	89%	55%
19	11,14	0,25	Computer Science	+	89%	59%
20	11,18	0,25	Computer Science	+	90%	62%
21	11,27	0,25	Computer Science	+	90%	66%
22	11,18	0,2	Business and Economics	-	86%	66%
23	10,49	0	Computer Science	+	87%	69%
24	10,72	0	Computer Science	+	88%	72%
25	10,72	0	Professional and Applied Computing	-	84%	72%
26	10,82	0	Professional and Applied Computing	-	81%	72%
27	10,95	0	Computer Science	+	81%	76%
28	11,00	0	Computer Science	+	82%	79%
29	11,00	0	Business and Economics	-	79%	79%
30	11,05	0	Medicine	-	77%	79%
31	11,05	0	Computer Science	+	77%	83%
32	11,09	0	Computer Science	+	78%	86%
33	11,14	0	Professional and Applied Computing	-	76%	86%
34	11,14	0	Computer Science	+	76%	90%
35	11,14	0	Computer Science	+	77%	93%
36	11,18	0	Business and Economics	-	75%	93%
37	11,18	0	Business and Economics	-	73%	93%
38	11,18	0	Medicine	-	71%	93%
39	11,18	0	Medicine	-	69%	93%
40	11,18	0	Medicine	-	68%	93%
41	11,18	0	Medicine	-	66%	93%
42	11,22	0	Business and Economics	-	64%	93%
43	11,22	0	Business and Economics	-	63%	93%
44	11,22	0	Biomedical and Life Sciences	-	61%	93%
45	11,22	0	Medicine	-	60%	93%
46	11,22	0	Computer Science	+	61%	97%
47	11,22	0	Computer Science	+	62%	100%
48	11,22	0	Medicine	-	60%	100%
49	11,27	0	Business and Economics	-	59%	100%
50	11,27	0	Medicine	-	58%	100%

Foram encontradas 29 (vinte e nove) publicações sobre o assunto "*Computer Science*" e gráfico da figura 4.6 representa a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve um ganho significativo, uma vez que inicialmente se tinha Precisão máxima (100%) até o 3º documento e a partir da reclassificação por IAE, este índice se manteve máximo até o 16º documento. A Recuperação que atingia o índice máximo apenas no 49º documento passou a atingir a recuperação máxima no 47º documento.

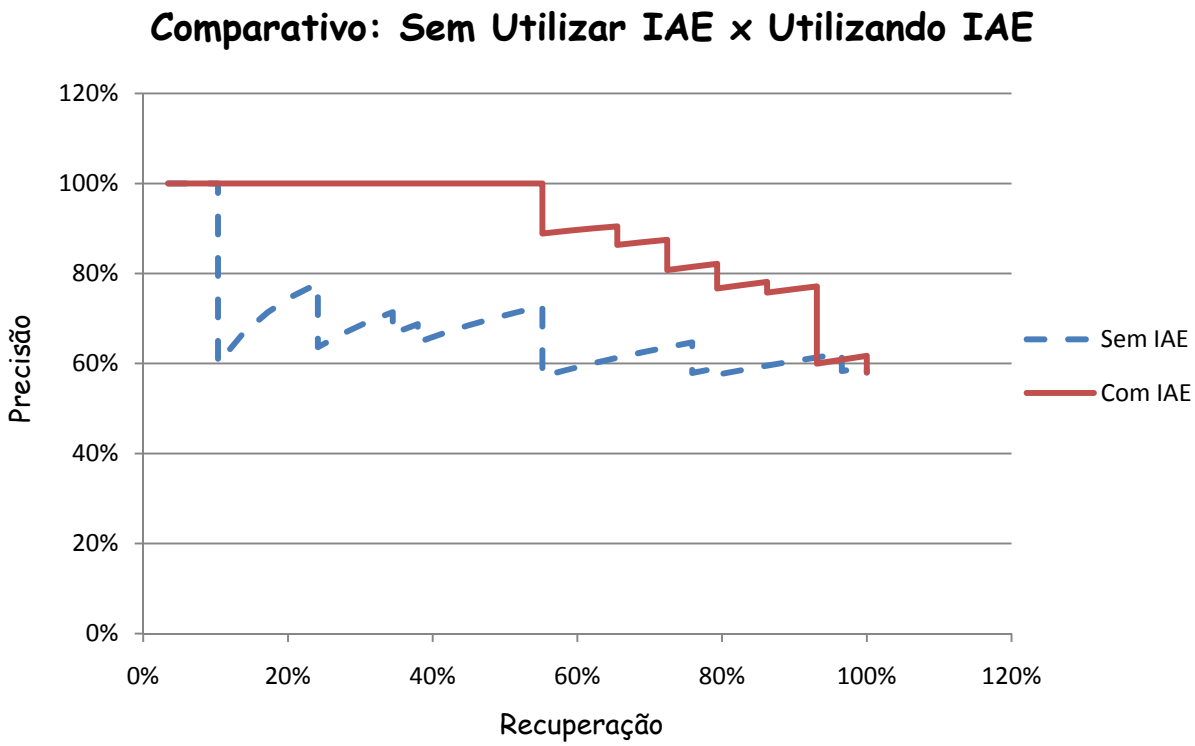


Figura 4.6 – Perfil 2: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma enorme melhoria na Precisão e grande melhoria na Recuperação de documentos.

4.3 Perfil 3 – Assunto: *Biomedical and Life Sciences*

Para o terceiro perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Biomedical and Life Sciences*". A tabela 4.9 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.10 e 4.11 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Biomedical and Life Sciences*" e negativo para os demais casos nas tabelas 4.10 e 4.11.

Tabela 4.9 – Perfil 3 utilizado na pesquisa. Assunto: "*Biomedical and Life Sciences*".

Assunto: ***Biomedical and Life Science***	
Título: ***A geography of ecosystem vulnerability***	
Palavras Relevantes:	**Peso**
1ª) *Biomedical*	50%
2ª) *Life*	25%
3ª) *System*	15%
4ª) *Sciences*	10%
Resumo: Land-cover change and the subsequent potential loss of natural resources due to conversion to anthropogenic use is regarded as one of the more pervasive environmental threats. Population and road data were used to generate interpolated surfaces of land demand across a large region, the mid-Atlantic states of Pennsylvania, Delaware, Maryland, Virginia, and West Virginia. The land demand surfaces were evaluated against land-cover change, as estimated using temporal decline in Normalized Difference Vegetation Index (NDVI). In general, the interpolated surfaces exhibited a plateau along the eastern seaboard that sank to a valley in the center of the study area, and then rose again to a plateau in the west that was of overall lower height than the plateau on the eastern seaboard. The spatial pattern of land-cover change showed the same general pattern as the interpolated surfaces of land demand. Correlations were significant regardless of variations used to generate the interpolated surfaces. The results suggest that human activity is the principal agent of land-cover change at regional scales in this region, and that natural resources that change as land cover changes (e.g., water, habitat) are exposed to a gradient of vulnerability that increases from west to east.	

Tabela 4.10 – Perfil 3, classificação por *KNN*. Assunto: "*Biomedical and Life Sciences*".

Doc.	*Distância*	*Asssunto*	*Relevância*	*Precisão*	*Recup.*
1	16,76	Business and Economics	-	0%	0%
2	17,03	Medicine	-	0%	0%
3	17,20	Computer Science	-	0%	0%
4	17,20	Computer Science	-	0%	0%
5	17,26	Computer Science	-	0%	0%
6	17,26	Medicine	-	0%	0%
7	17,32	Computer Science	-	0%	0%
8	17,35	Computer Science	-	0%	0%
9	17,35	Computer Science	-	0%	0%
10	17,38	Computer Science	-	0%	0%
11	17,38	Computer Science	-	0%	0%
12	17,46	Business and Economics	-	0%	0%
13	17,46	Computer Science	-	0%	0%
14	17,46	Computer Science	-	0%	0%
15	17,49	Computer Science	-	0%	0%
16	17,55	Architecture and Design	-	0%	0%
17	17,58	Biomedical and Life Sciences	+	6%	33%
18	17,58	Computer Science	-	6%	33%
19	17,61	Professional and Applied Computing	-	5%	33%
20	17,61	Professional and Applied Computing	-	5%	33%
21	17,64	Computer Science	-	5%	33%
22	17,64	Computer Science	-	5%	33%
23	17,66	Architecture and Design	-	4%	33%
24	17,66	Business and Economics	-	4%	33%
25	17,66	Biomedical and Life Sciences	+	8%	67%
26	17,69	Computer Science	-	8%	67%
27	17,69	Architecture and Design	-	7%	67%
28	17,69	Computer Science	-	7%	67%
29	17,72	Business and Economics	-	7%	67%
30	17,72	Biomedical and Life Sciences	+	10%	100%
31	17,75	Computer Science	-	10%	100%
32	17,75	Professional and Applied Computing	-	9%	100%
33	17,75	Professional and Applied Computing	-	9%	100%
34	17,78	Architecture and Design	-	9%	100%
35	17,78	Computer Science	-	9%	100%
36	17,78	Computer Science	-	8%	100%
37	17,78	Business and Economics	-	8%	100%
38	17,78	Computer Science	-	8%	100%
39	17,80	Professional and Applied Computing	-	8%	100%
40	17,80	Business and Economics	-	8%	100%
41	17,80	Computer Science	-	7%	100%
42	17,83	Professional and Applied Computing	-	7%	100%
43	17,83	Professional and Applied Computing	-	7%	100%
44	17,83	Business and Economics	-	7%	100%
45	17,86	Business and Economics	-	7%	100%
46	17,86	Business and Economics	-	7%	100%
47	17,86	Computer Science	-	6%	100%
48	17,86	Professional and Applied Computing	-	6%	100%
49	17,86	Computer Science	-	6%	100%
50	17,89	Business and Economics	-	6%	100%

Tabela 4.11 – Perfil 3, reclassificação por IAE. Assunto: "*Biomedical and Life Sciences*".

Doc.	*Distância*	*IAE*	*Assunto*	*Relevância*	*Precisão*	*Recup.*
1	17,83	0,35	Professional and Applied Computing	-	0%	0%
2	17,26	0,3	Computer Science	-	0%	0%
3	17,66	0,15	Biomedical and Life Sciences	+	33%	33%
4	17,83	0,1	Professional and Applied Computing	-	25%	33%
5	16,76	0	Business and Economics	-	20%	33%
6	17,03	0	Medicine	-	17%	33%
7	17,20	0	Computer Science	-	14%	33%
8	17,20	0	Computer Science	-	13%	33%
9	17,26	0	Medicine	-	11%	33%
10	17,32	0	Computer Science	-	10%	33%
11	17,35	0	Computer Science	-	9%	33%
12	17,35	0	Computer Science	-	8%	33%
13	17,38	0	Computer Science	-	8%	33%
14	17,38	0	Computer Science	-	7%	33%
15	17,46	0	Business and Economics	-	7%	33%
16	17,46	0	Computer Science	-	6%	33%
17	17,46	0	Computer Science	-	6%	33%
18	17,49	0	Computer Science	-	6%	33%
19	17,55	0	Architecture and Design	-	5%	33%
20	17,58	0	Biomedical and Life Sciences	+	10%	67%
21	17,58	0	Computer Science	-	10%	67%
22	17,61	0	Professional and Applied Computing	-	9%	67%
23	17,61	0	Professional and Applied Computing	-	9%	67%
24	17,64	0	Computer Science	-	8%	67%
25	17,64	0	Computer Science	-	8%	67%
26	17,66	0	Architecture and Design	-	8%	67%
27	17,66	0	Business and Economics	-	7%	67%
28	17,69	0	Computer Science	-	7%	67%
29	17,69	0	Architecture and Design	-	7%	67%
30	17,69	0	Computer Science	-	7%	67%
31	17,72	0	Business and Economics	-	6%	67%
32	17,72	0	Biomedical and Life Sciences	+	9%	100%
33	17,75	0	Computer Science	-	9%	100%
34	17,75	0	Professional and Applied Computing	-	9%	100%
35	17,75	0	Professional and Applied Computing	-	9%	100%
36	17,78	0	Architecture and Design	-	8%	100%
37	17,78	0	Computer Science	-	8%	100%
38	17,78	0	Computer Science	-	8%	100%
39	17,78	0	Business and Economics	-	8%	100%
40	17,78	0	Computer Science	-	8%	100%
41	17,80	0	Professional and Applied Computing	-	7%	100%
42	17,80	0	Business and Economics	-	7%	100%
43	17,80	0	Computer Science	-	7%	100%
44	17,83	0	Business and Economics	-	7%	100%
45	17,86	0	Business and Economics	-	7%	100%
46	17,86	0	Business and Economics	-	7%	100%
47	17,86	0	Computer Science	-	6%	100%
48	17,86	0	Professional and Applied Computing	-	6%	100%
49	17,86	0	Computer Science	-	6%	100%
50	17,89	0	Business and Economics	-	6%	100%

Foram encontradas 3 (três) publicações sobre o assunto: "*Biomedical and Life Sciences*" e o gráfico da figura 4.7 representa a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve um ganho significativo, uma vez que inicialmente se tinha Precisão de 6% somente no 17º documento e a partir da reclassificação por IAE, este índice se passou para 33% no 3º documento. A Recuperação que atingia o índice máximo no 30º documento passou a atingir a recuperação máxima no 32º documento. Houve perda de Recuperação, que pode ser melhorado com uma melhor escolha das palavras relevantes ao assunto pesquisado.

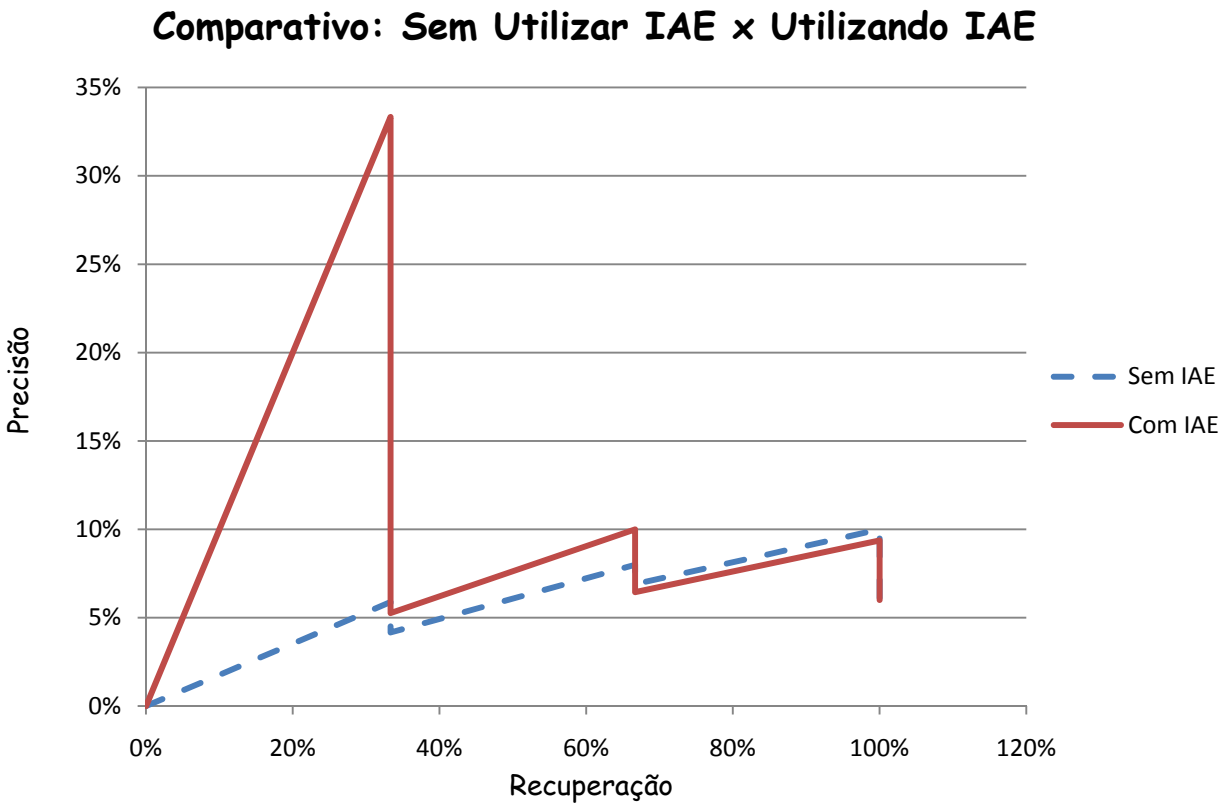


Figura 4.7 – Perfil 3: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma substancial melhoria na Precisão, mas ocorreu uma ligeira piora ao atingir o percentual máximo na Recuperação de documentos.

4.4 Perfil 4 – Assunto: *Behavioral Science*

Para o quarto perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Behavioral Science*". A tabela 4.12 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.13 e 4.14 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Behavioral Science*" e negativo para os demais casos nas tabelas 4.13 e 4.14.

Tabela 4.12 – Perfil 4 utilizado na pesquisa. Assunto: "*Behavioral Science*".

Assunto: ***Behavioral Science***	
Título: ***A study in word formation restrictions***	
Palavras Relevantes:	**Peso**
1ª) *Behavioral*	50%
2ª) *Substance*	25%
3ª) *Function*	15%
4ª) *Science*	10%
Resumo: Subjects judged the acceptability of pseudo words formed a stem, an optional affix, and-ity. The results show that those words in which-ity was adjacent to a latinate morpheme were more acceptable than those in which it was not. Other factors which influenced the acceptability ratings were familiarity and homogeneity of morpheme types. In their evaluations of the possible words, subjects were sensitive to historical class distinctions of stems and affixes.	

Tabela 4.13 – Perfil 4, classificação por *KNN*. Assunto: "*Behavioral Science*".

Doc.	Distância	Assunto	Relevância	Precisão	Recup.
1	8,37	Medicine	-	0%	0%
2	8,43	Medicine	-	0%	0%
3	8,43	Business and Economics	-	0%	0%
4	8,54	Computer Science	-	0%	0%
5	8,60	Computer Science	-	0%	0%
6	8,66	Computer Science	-	0%	0%
7	8,72	Computer Science	-	0%	0%
8	8,72	Architecture and Design	-	0%	0%
9	8,72	Architecture and Design	-	0%	0%
10	8,72	Architecture and Design	-	0%	0%
11	8,72	Medicine	-	0%	0%
12	8,72	Medicine	-	0%	0%
13	8,72	Medicine	-	0%	0%
14	8,72	Medicine	-	0%	0%
15	8,72	Business and Economics	-	0%	0%
16	8,72	Business and Economics	-	0%	0%
17	8,72	Business and Economics	-	0%	0%
18	8,72	Business and Economics	-	0%	0%
19	8,72	Biomedical and Life Sciences	-	0%	0%
20	8,72	Biomedical and Life Sciences	-	0%	0%
21	8,72	Biomedical and Life Sciences	-	0%	0%
22	8,77	Biomedical and Life Sciences	-	0%	0%
23	8,77	Biomedical and Life Sciences	-	0%	0%
24	8,77	Biomedical and Life Sciences	-	0%	0%
25	8,77	Biomedical and Life Sciences	-	0%	0%
26	8,77	Biomedical and Life Sciences	-	0%	0%
27	8,77	Business and Economics	-	0%	0%
28	8,77	Business and Economics	-	0%	0%
29	8,77	Business and Economics	-	0%	0%
30	8,77	Business and Economics	-	0%	0%
31	8,77	Computer Science	-	0%	0%
32	8,77	Computer Science	-	0%	0%
33	8,77	Medicine	-	0%	0%
34	8,77	Medicine	-	0%	0%
35	8,77	Medicine	-	0%	0%
36	8,77	Medicine	-	0%	0%
37	8,77	Medicine	-	0%	0%
38	8,77	Biomedical and Life Sciences	-	0%	0%
39	8,83	Computer Science	-	0%	0%
40	8,83	Medicine	-	0%	0%
41	8,83	Computer Science	-	0%	0%
42	8,83	Computer Science	-	0%	0%
43	8,83	Professional and Applied Computing	-	0%	0%
44	8,83	Business and Economics	-	0%	0%
45	8,83	Business and Economics	-	0%	0%
46	8,83	Biomedical and Life Sciences	-	0%	0%
47	8,83	Computer Science	-	0%	0%
48	8,83	Medicine	-	0%	0%
49	8,83	Business and Economics	-	0%	0%
50	8,83	Biomedical and Life Sciences	-	0%	0%

Tabela 4.14 – Perfil 4, reclassificação por IAE. Assunto: "*Behavioral Science*".

Doc.	Distância	IAE	Asssunto	Relevância	Precisão	Recup.
1	8,66	1	Computer Science	-	0%	0%
2	8,83	0,5	Computer Science	-	0%	0%
3	8,72	0,15	Biomedical and Life Sciences	-	0%	0%
4	8,72	0,15	Biomedical and Life Sciences	-	0%	0%
5	8,77	0,15	Medicine	-	0%	0%
6	8,83	0,15	Computer Science	-	0%	0%
7	8,83	0,1	Biomedical and Life Sciences	-	0%	0%
8	8,37	0	Medicine	-	0%	0%
9	8,43	0	Medicine	-	0%	0%
10	8,43	0	Business and Economics	-	0%	0%
11	8,54	0	Computer Science	-	0%	0%
12	8,60	0	Computer Science	-	0%	0%
13	8,72	0	Computer Science	-	0%	0%
14	8,72	0	Architecture and Design	-	0%	0%
15	8,72	0	Architecture and Design	-	0%	0%
16	8,72	0	Architecture and Design	-	0%	0%
17	8,72	0	Medicine	-	0%	0%
18	8,72	0	Medicine	-	0%	0%
19	8,72	0	Medicine	-	0%	0%
20	8,72	0	Medicine	-	0%	0%
21	8,72	0	Business and Economics	-	0%	0%
22	8,72	0	Business and Economics	-	0%	0%
23	8,72	0	Business and Economics	-	0%	0%
24	8,72	0	Business and Economics	-	0%	0%
25	8,72	0	Biomedical and Life Sciences	-	0%	0%
26	8,77	0	Biomedical and Life Sciences	-	0%	0%
27	8,77	0	Biomedical and Life Sciences	-	0%	0%
28	8,77	0	Biomedical and Life Sciences	-	0%	0%
29	8,77	0	Biomedical and Life Sciences	-	0%	0%
30	8,77	0	Biomedical and Life Sciences	-	0%	0%
31	8,77	0	Business and Economics	-	0%	0%
32	8,77	0	Business and Economics	-	0%	0%
33	8,77	0	Business and Economics	-	0%	0%
34	8,77	0	Business and Economics	-	0%	0%
35	8,77	0	Computer Science	-	0%	0%
36	8,77	0	Computer Science	-	0%	0%
37	8,77	0	Medicine	-	0%	0%
38	8,77	0	Medicine	-	0%	0%
39	8,77	0	Medicine	-	0%	0%
40	8,77	0	Medicine	-	0%	0%
41	8,77	0	Biomedical and Life Sciences	-	0%	0%
42	8,83	0	Medicine	-	0%	0%
43	8,83	0	Computer Science	-	0%	0%
44	8,83	0	Professional and Applied Computing	-	0%	0%
45	8,83	0	Business and Economics	-	0%	0%
46	8,83	0	Business and Economics	-	0%	0%
47	8,83	0	Computer Science	-	0%	0%
48	8,83	0	Medicine	-	0%	0%
49	8,83	0	Business and Economics	-	0%	0%
50	8,83	0	Biomedical and Life Sciences	-	0%	0%

Como era esperado, não foram encontradas publicações sobre o assunto: "*Behavioral Science*" por falta de documentos correlacionados a este assunto com número mínimo de palavras relevantes de 25 palavras. A Precisão e a Recuperação ficaram com índice 0 (zero), não podendo ser levada em consideração para a análise de resultados. Uma melhor definição no número mínimo de palavras relevantes em cada publicação pode permitir um resultado melhor.

4.5 Perfil 5 – Assunto: *Business and Economics*

Para o quinto perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Business and Economics*". A tabela 4.15 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.16 e 4.17 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Business and Economics*" e negativo para os demais casos nas tabelas 4.16 e 4.17.

Tabela 4.15 – Perfil 5 utilizado na pesquisa. Assunto: "*Business and Economics*".

Assunto: ***Business and Economics***	
Título: ***Analyzing Mathematical Programs Using MProbe***	
Palavras Relevantes:	**Peso**
1ª) *Business*	50%
2ª) *Economics*	25%
3ª) *Management*	15%
4ª) *Global*	10%
Resumo: Just as modern general-purpose programming languages (e.g., C++, Java) are supported by a suite of tools (debuggers, profilers, etc.), mathematical programming languages need supporting tools. MProbe is an example of a suite of tools supporting a mathematical programming language, in this case AMPL. MProbe includes tools for empirically estimating the shape of nonlinear functions of many variables, nonlinearly-constrained region shape, the effect of the objective shape on the ability to find a global optimum, tools for estimating the effectiveness of constraints and for navigating through the model, among others.	

Tabela 4.16 – Perfil 5, classificação por *KNN*. Assunto: "*Business and Economics*".

Doc.	Distância	Assunto	Relevância	Precisão	Recup.
1	9,80	Computer Science	-	0%	0%
2	10,00	Medicine	-	0%	0%
3	10,10	Medicine	-	0%	0%
4	10,15	Computer Science	-	0%	0%
5	10,25	Computer Science	-	0%	0%
6	10,30	Business and Economics	+	17%	25%
7	10,30	Computer Science	-	14%	25%
8	10,34	Business and Economics	+	25%	50%
9	10,34	Computer Science	-	22%	50%
10	10,39	Computer Science	-	20%	50%
11	10,44	Computer Science	-	18%	50%
12	10,44	Business and Economics	+	25%	75%
13	10,49	Computer Science	-	23%	75%
14	10,49	Professional and Applied Computing	-	21%	75%
15	10,54	Medicine	-	20%	75%
16	10,54	Computer Science	-	19%	75%
17	10,58	Medicine	-	18%	75%
18	10,58	Computer Science	-	17%	75%
19	10,58	Computer Science	-	16%	75%
20	10,58	Medicine	-	15%	75%
21	10,63	Architecture and Design	-	14%	75%
22	10,68	Medicine	-	14%	75%
23	10,68	Architecture and Design	-	13%	75%
24	10,68	Computer Science	-	13%	75%
25	10,72	Computer Science	-	12%	75%
26	10,72	Computer Science	-	12%	75%
27	10,72	Professional and Applied Computing	-	11%	75%
28	10,72	Computer Science	-	11%	75%
29	10,72	Computer Science	-	10%	75%
30	10,72	Computer Science	-	10%	75%
31	10,77	Computer Science	-	10%	75%
32	10,77	Architecture and Design	-	9%	75%
33	10,77	Computer Science	-	9%	75%
34	10,82	Computer Science	-	9%	75%
35	10,82	Biomedical and Life Sciences	-	9%	75%
36	10,82	Computer Science	-	8%	75%
37	10,86	Computer Science	-	8%	75%
38	10,86	Computer Science	-	8%	75%
39	10,91	Professional and Applied Computing	-	8%	75%
40	10,91	Biomedical and Life Sciences	-	8%	75%
41	10,91	Biomedical and Life Sciences	-	7%	75%
42	10,91	Computer Science	-	7%	75%
43	10,91	Computer Science	-	7%	75%
44	10,95	Computer Science	-	7%	75%
45	10,95	Computer Science	-	7%	75%
46	11,00	Computer Science	-	7%	75%
47	11,05	Business and Economics	+	9%	100%
48	11,05	Biomedical and Life Sciences	-	8%	100%
49	11,05	Computer Science	-	8%	100%
50	11,05	Professional and Applied Computing	-	8%	100%

Tabela 4.17 – Perfil 5, reclassificação por IAE. Assunto: "*Business and Economics*".

Doc.	Distância	IAE	Assunto	Relevância	Precisão	Recup.
1	10,72	0,15	Computer Science	-	0%	0%
2	9,80	0	Computer Science	-	0%	0%
3	10,00	0	Medicine	-	0%	0%
4	10,10	0	Medicine	-	0%	0%
5	10,15	0	Computer Science	-	0%	0%
6	10,25	0	Computer Science	-	0%	0%
7	10,30	0	Business and Economics	+	14%	25%
8	10,30	0	Computer Science	-	13%	25%
9	10,34	0	Business and Economics	+	22%	50%
10	10,34	0	Computer Science	-	20%	50%
11	10,39	0	Computer Science	-	18%	50%
12	10,44	0	Computer Science	-	17%	50%
13	10,44	0	Business and Economics	+	23%	75%
14	10,49	0	Computer Science	-	21%	75%
15	10,49	0	Professional and Applied Computing	-	20%	75%
16	10,54	0	Medicine	-	19%	75%
17	10,54	0	Computer Science	-	18%	75%
18	10,58	0	Medicine	-	17%	75%
19	10,58	0	Computer Science	-	16%	75%
20	10,58	0	Computer Science	-	15%	75%
21	10,58	0	Medicine	-	14%	75%
22	10,63	0	Architecture and Design	-	14%	75%
23	10,68	0	Medicine	-	13%	75%
24	10,68	0	Architecture and Design	-	13%	75%
25	10,68	0	Computer Science	-	12%	75%
26	10,72	0	Computer Science	-	12%	75%
27	10,72	0	Computer Science	-	11%	75%
28	10,72	0	Professional and Applied Computing	-	11%	75%
29	10,72	0	Computer Science	-	10%	75%
30	10,72	0	Computer Science	-	10%	75%
31	10,77	0	Computer Science	-	10%	75%
32	10,77	0	Architecture and Design	-	9%	75%
33	10,77	0	Computer Science	-	9%	75%
34	10,82	0	Computer Science	-	9%	75%
35	10,82	0	Biomedical and Life Sciences	-	9%	75%
36	10,82	0	Computer Science	-	8%	75%
37	10,86	0	Computer Science	-	8%	75%
38	10,86	0	Computer Science	-	8%	75%
39	10,91	0	Professional and Applied Computing	-	8%	75%
40	10,91	0	Biomedical and Life Sciences	-	8%	75%
41	10,91	0	Biomedical and Life Sciences	-	7%	75%
42	10,91	0	Computer Science	-	7%	75%
43	10,91	0	Computer Science	-	7%	75%
44	10,95	0	Computer Science	-	7%	75%
45	10,95	0	Computer Science	-	7%	75%
46	11,00	0	Computer Science	-	7%	75%
47	11,05	0	Business and Economics	+	9%	100%
48	11,05	0	Biomedical and Life Sciences	-	8%	100%
49	11,05	0	Computer Science	-	8%	100%
50	11,05	0	Professional and Applied Computing	-	8%	100%

Foram encontradas 4 (quatro) publicações sobre o assunto: "*Business and Economics*" e o gráfico da figura 4.8 representa a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve uma pequena perda, uma vez que inicialmente se tinha Precisão de 17% no 6º documento e a partir da reclassificação por IAE, este índice se passou para 14% no 7º documento. A Recuperação que atingia o índice máximo no 47º documento se manteve. Houve perda de Precisão, que pode ser melhorado com uma melhor escolha das palavras relevantes ao assunto pesquisado.

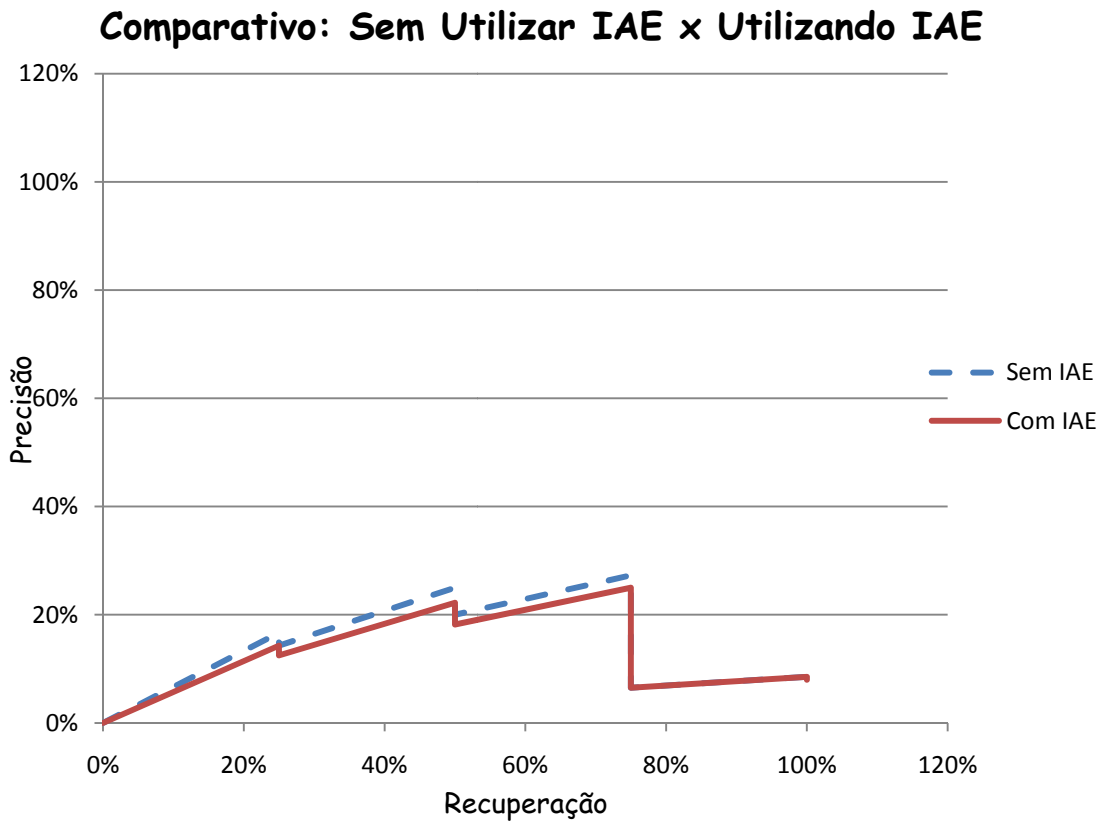


Figura 4.8 – Perfil 5: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma ligeira piora, mas que não compromete o resultado alcançado, uma vez que apesar de reposicionar o primeiro documento encontrado uma posição abaixo, ele eleva em uma posição as duas próximas publicações relevantes ao tema. Foi observado que o uso do IAE tem grande melhoria quando usado em conjunto com resumos com maior número de palavras relevantes. O escolha dos termos mais relevantes para emprego no IAE não foi satisfatório.

4.6 Perfil 6 – Assunto: *Architeture and Design*

Para o sexto perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Architeture and Design*". A tabela 4.18 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.19 e 4.20 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Architeture and Design*" e negativo para os demais casos nas tabelas 4.18 e 4.20.

Tabela 4.18 – Perfil 6 utilizado na pesquisa. Assunto: "*Architeture and Design*".

Assunto: ***Architetura and Design***	
Título: ***Maps and Geopolitics in Video Games***	
Palavras Relevantes:	**Peso**
1ª) *Design*	50%
2ª) *Street*	25%
3ª) *Place*	15%
4ª) *Park*	10%
Resumo: With the introduction of 3D-graphics space, a fundamental change in video game aesthetics and performativity took place: virtual space is now predominantly generated as presented space or spatial presence (Poole 2000/McMahan 2003). However, spatial formations in video games confusingly are still referred to as "representational space" (Wolf 1997) in general, even though the pictorial appearance of the game is not always a literal representation. Rather, representations in video games have specific functions that define their role in the interactive play just like any particular form of spatiality does (Aarseth 2001/Fernández-Vara et al. 2005/Taylor 2005). The function of representations in games can sufficiently be described by Henri Lefebvre's (1991) trialectic of spatial processes, according to which the individual's perception of spacetime and the spatiotemporal structures of the social — "representational spaces" in Lefebvre's words — are reciprocally mediated by representations of space — namely, maps. Space in Lefebvre's understanding is thus threefold: a combination of perceived (perçu), conceived (conçu) and lived (vécu). Video games today are mostly a presentation of perceptual space in the way Lefebvre addresses the individual experience of space or what he also calls the "spatial practice." In contrast, representations of space differ from this phenomenal experience of space: as they are in real-life contexts, video game maps are essential for orientation, especially in games played from the first-person perspective, for in those games, one not only needs to see what one is aiming at, but also where one is located within the entire setting of the game. For this reason, maps in video games are either fully displayed and function as representations of the whole "playground" (which gamers usually call the "map"), or they are reduced to a visual element within the display, most frequently a radar that allows for orientation within the periphery of the position of the avatar or the ego in play.	

Tabela 4.19 – Perfil 6, classificação por *KNN*. Assunto: "*Architeture and Design*".

Doc.	Distância	Assunto	Relevância	Precisão	Recup.
1	2,83	Architecture and Design	+	100%	8%
2	18,92	Business and Economics	-	50%	8%
3	19,21	Architecture and Design	+	67%	15%
4	19,44	Computer Science	-	50%	15%
5	20,10	Computer Science	-	40%	15%
6	20,10	Computer Science	-	33%	15%
7	20,12	Computer Science	-	29%	15%
8	20,27	Architecture and Design	+	38%	23%
9	20,27	Computer Science	-	33%	23%
10	20,37	Computer Science	-	30%	23%
11	20,37	Computer Science	-	27%	23%
12	20,40	Architecture and Design	+	33%	31%
13	20,40	Computer Science	-	31%	31%
14	20,42	Computer Science	-	29%	31%
15	20,45	Computer Science	-	27%	31%
16	20,47	Computer Science	-	25%	31%
17	20,52	Professional and Applied Computing	-	24%	31%
18	20,62	Computer Science	-	22%	31%
19	20,71	Computer Science	-	21%	31%
20	20,76	Biomedical and Life Sciences	-	20%	31%
21	20,88	Business and Economics	-	19%	31%
22	21,42	Business and Economics	-	18%	31%
23	21,52	Architecture and Design	+	22%	38%
24	21,61	Architecture and Design	+	25%	46%
25	21,61	Professional and Applied Computing	-	24%	46%
26	21,70	Architecture and Design	+	27%	54%
27	21,75	Computer Science	-	26%	54%
28	22,00	Professional and Applied Computing	-	25%	54%
29	22,00	Computer Science	-	24%	54%
30	22,00	Computer Science	-	23%	54%
31	22,00	Professional and Applied Computing	-	23%	54%
32	22,05	Computer Science	-	22%	54%
33	22,05	Computer Science	-	21%	54%
34	22,11	Professional and Applied Computing	-	21%	54%
35	22,18	Architecture and Design	+	23%	62%
36	22,20	Architecture and Design	+	25%	69%
37	22,23	Computer Science	-	24%	69%
38	22,23	Computer Science	-	24%	69%
39	22,34	Computer Science	-	23%	69%
40	22,43	Professional and Applied Computing	-	23%	69%
41	22,45	Architecture and Design	+	24%	77%
42	22,45	Business and Economics	-	24%	77%
43	22,47	Business and Economics	-	23%	77%
44	22,49	Professional and Applied Computing	-	23%	77%
45	22,49	Professional and Applied Computing	-	22%	77%
46	22,49	Professional and Applied Computing	-	22%	77%
47	22,52	Architecture and Design	+	23%	85%
48	22,54	Architecture and Design	+	25%	92%
49	22,54	Business and Economics	-	24%	92%
50	22,58	Architecture and Design	+	26%	100%

Tabela 4.20 – Perfil 6, reclassificação por IAE. Assunto: "*Architeture and Design*".

Doc.	Distância	IAE	Assunto	Relevância	Precisão	Recup.
1	22,05	1,5	Computer Science	-	0%	0%
2	22,05	1,5	Computer Science	-	0%	0%
3	22,11	1,5	Professional and Applied Computing	-	0%	0%
4	20,40	1	Architecture and Design	+	25%	8%
5	20,62	1	Computer Science	-	20%	8%
6	22,00	1	Computer Science	-	17%	8%
7	22,45	1	Architecture and Design	+	29%	15%
8	21,61	0,65	Architecture and Design	+	38%	23%
9	20,37	0,5	Computer Science	-	33%	23%
10	20,40	0,5	Computer Science	-	30%	23%
11	20,52	0,5	Professional and Applied Computing	-	27%	23%
12	21,61	0,5	Professional and Applied Computing	-	25%	23%
13	22,18	0,5	Architecture and Design	+	31%	31%
14	21,52	0,3	Architecture and Design	+	36%	38%
15	2,83	0,15	Architecture and Design	+	40%	46%
16	19,21	0,15	Architecture and Design	+	44%	54%
17	18,92	0	Business and Economics	-	41%	54%
18	19,44	0	Computer Science	-	39%	54%
19	20,10	0	Computer Science	-	37%	54%
20	20,10	0	Computer Science	-	35%	54%
21	20,12	0	Computer Science	-	33%	54%
22	20,27	0	Architecture and Design	+	36%	62%
23	20,27	0	Computer Science	-	35%	62%
24	20,37	0	Computer Science	-	33%	62%
25	20,42	0	Computer Science	-	32%	62%
26	20,45	0	Computer Science	-	31%	62%
27	20,47	0	Computer Science	-	30%	62%
28	20,71	0	Computer Science	-	29%	62%
29	20,76	0	Biomedical and Life Sciences	-	28%	62%
30	20,88	0	Business and Economics	-	27%	62%
31	21,42	0	Business and Economics	-	26%	62%
32	21,70	0	Architecture and Design	+	28%	69%
33	21,75	0	Computer Science	-	27%	69%
34	22,00	0	Professional and Applied Computing	-	26%	69%
35	22,00	0	Computer Science	-	26%	69%
36	22,00	0	Professional and Applied Computing	-	25%	69%
37	22,20	0	Architecture and Design	+	27%	77%
38	22,23	0	Computer Science	-	26%	77%
39	22,23	0	Computer Science	-	26%	77%
40	22,34	0	Computer Science	-	25%	77%
41	22,43	0	Professional and Applied Computing	-	24%	77%
42	22,45	0	Business and Economics	-	24%	77%
43	22,47	0	Business and Economics	-	23%	77%
44	22,49	0	Professional and Applied Computing	-	23%	77%
45	22,49	0	Professional and Applied Computing	-	22%	77%
46	22,49	0	Professional and Applied Computing	-	22%	77%
47	22,52	0	Architecture and Design	+	23%	85%
48	22,54	0	Architecture and Design	+	25%	92%
49	22,54	0	Business and Economics	-	24%	92%
50	22,58	0	Architecture and Design	+	26%	100%

Foram encontradas 13 (treze) publicações sobre o assunto: "*Architeture and Design*" e o gráfico da figura 4.9 apresenta a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve uma perda significativa, uma vez que inicialmente se tinha Precisão de 67% no 3º documento e a partir da reclassificação por IAE, este índice se passou para 25% no 4º documento. A Recuperação que atingia o índice máximo no 50º documento se manteve. Houve perda de Precisão, que pode ser melhorado com uma melhor escolha das palavras relevantes ao assunto pesquisado.

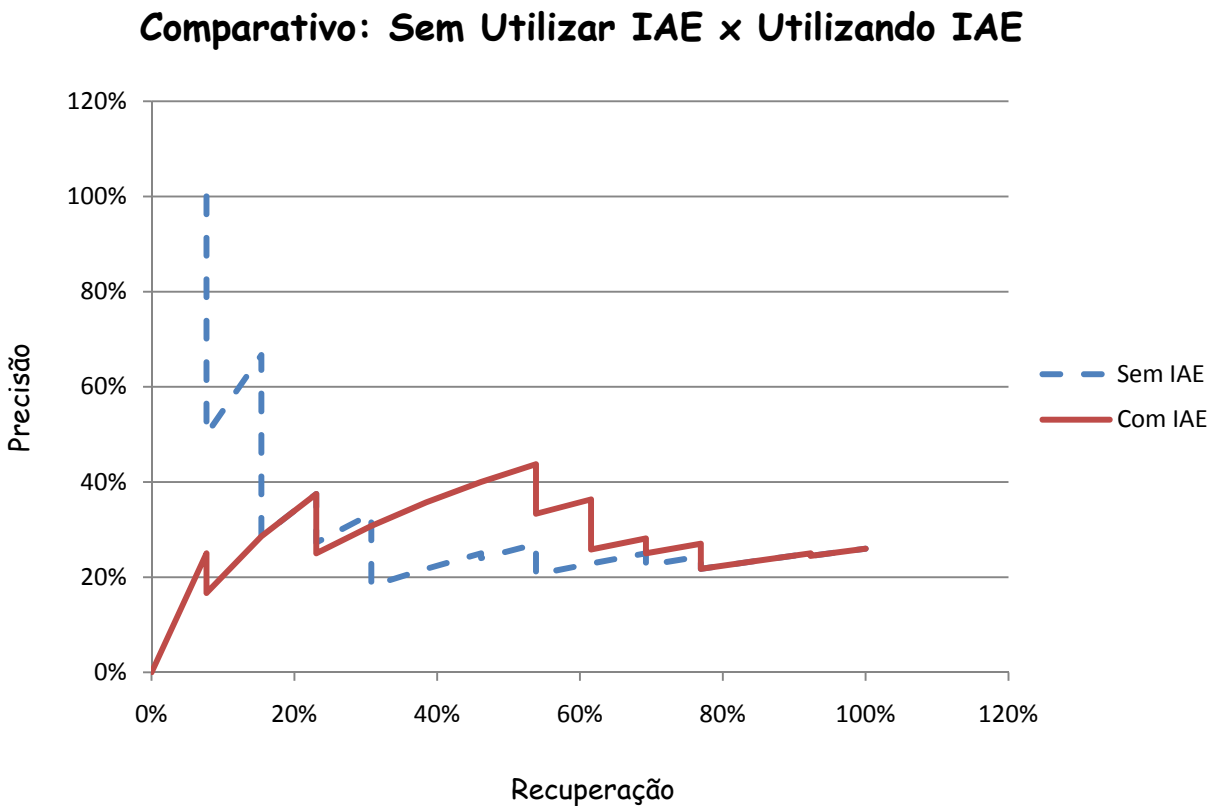


Figura 4.9 – Perfil 6: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma piora inicial na Precisão, mas que melhora ao decorrer das demais recuperações. Como o resumo do perfil não é de poucas palavras relevantes, Concluí-se que a escolha dos termos mais relevantes para este assunto foi mal escolhido.

4.7 Perfil 7 – Assunto: *Professional and Applied Computing*

Para o sétimo perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Professional and Applied Computing*". A tabela 4.21 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.22 e 4.23 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Professional and Applied Computing*" e negativo para os demais casos nas tabelas 4.22 e 4.23.

Tabela 4.21 – Perfil 7 utilizado na pesquisa. Assunto: *"Professional and Applied Computing*".

Assunto: ***Professional and Applied Computing***	
Título: ***Zone Files and Resource Records***	
Palavras Relevantes:	**Peso**
1ª) *Draw*	50%
2ª) *Graphics*	25%
3ª) *Animation*	15%
4ª) *Process*	10%
Resumo: This chapter described the format and content of zone files. You learned about the $TTL directive, used to set the default TTL for the zone. You also encountered the $ORIGIN directive, used to set the base name for the zone, and the $ORIGIN substitution rule, the cause of much DNS aggravation. Using the example zone file as a guide, the text explained the various Resource Record types used to construct basic zone files such as the Start of Authority, Name Server, Mail Exchanger, and Address Resource Records. Chapter 3 explains DNS operations: the types of DNS queries that may be used; reverse mapping, the process by which an IP address may be mapped to a host name; zone transfers, the method by which zone files are updated from the master to the slave name servers; and finally, a brief overview of the security issues involved in running a DNS service.	

Tabela 4.22 – Perfil 7, classificação por *KNN*. Assunto: "*Professional and Applied Computing*".

Doc.	Distância	Assunto	Relevância	Precisão	Recup.
1	2,24	Professional and Applied Computing	+	100%	8%
2	12,57	Professional and Applied Computing	+	100%	15%
3	12,61	Professional and Applied Computing	+	100%	23%
4	12,61	Computer Science	-	75%	23%
5	12,61	Medicine	-	60%	23%
6	12,61	Business and Economics	-	50%	23%
7	12,73	Medicine	-	43%	23%
8	12,73	Medicine	-	38%	23%
9	12,73	Business and Economics	-	33%	23%
10	12,77	Business and Economics	-	30%	23%
11	12,77	Computer Science	-	27%	23%
12	12,77	Computer Science	-	25%	23%
13	12,77	Medicine	-	23%	23%
14	12,81	Medicine	-	21%	23%
15	12,81	Architecture and Design	-	20%	23%
16	12,81	Computer Science	-	19%	23%
17	12,85	Computer Science	-	18%	23%
18	12,85	Medicine	-	17%	23%
19	12,85	Architecture and Design	-	16%	23%
20	12,85	Professional and Applied Computing	+	20%	31%
21	12,85	Professional and Applied Computing	+	24%	38%
22	12,88	Medicine	-	23%	38%
23	12,88	Computer Science	-	22%	38%
24	12,88	Computer Science	-	21%	38%
25	12,88	Professional and Applied Computing	+	24%	46%
26	12,88	Computer Science	-	23%	46%
27	12,92	Professional and Applied Computing	+	26%	54%
28	12,92	Professional and Applied Computing	+	29%	62%
29	12,92	Professional and Applied Computing	+	31%	69%
30	12,92	Computer Science	-	30%	69%
31	12,92	Computer Science	-	29%	69%
32	12,92	Computer Science	-	28%	69%
33	12,92	Architecture and Design	-	27%	69%
34	12,92	Architecture and Design	-	26%	69%
35	12,92	Business and Economics	-	26%	69%
36	12,92	Business and Economics	-	25%	69%
37	12,92	Business and Economics	-	24%	69%
38	12,96	Computer Science	-	24%	69%
39	12,96	Computer Science	-	23%	69%
40	12,96	Business and Economics	-	23%	69%
41	12,96	Computer Science	-	22%	69%
42	12,96	Medicine	-	21%	69%
43	12,96	Business and Economics	-	21%	69%
44	12,96	Computer Science	-	20%	69%
45	12,96	Professional and Applied Computing	+	22%	77%
46	12,96	Computer Science	-	22%	77%
47	13,00	Professional and Applied Computing	+	23%	85%
48	13,00	Computer Science	-	23%	85%
49	13,00	Professional and Applied Computing	+	24%	92%
50	13,00	Professional and Applied Computing	+	26%	100%

Tabela 4.23 – Perfil 7, reclassificação por IAE. Assunto: "*Professional and Applied Computing*".

Doc.	Distância	IAE	Assunto	Relevância	Precisão	Recup.
1	12,96	0,75	Architecture and Design	-	0%	0%
2	2,24	0,1	Professional and Applied Computing	+	50%	8%
3	12,57	0,1	Professional and Applied Computing	+	67%	17%
4	12,85	0,1	Professional and Applied Computing	+	75%	25%
5	12,61	0	Professional and Applied Computing	+	80%	33%
6	12,61	0	Computer Science	-	67%	33%
7	12,61	0	Medicine	-	57%	33%
8	12,61	0	Business and Economics	-	50%	33%
9	12,73	0	Medicine	-	44%	33%
10	12,73	0	Medicine	-	40%	33%
11	12,73	0	Business and Economics	-	36%	33%
12	12,77	0	Business and Economics	-	33%	33%
13	12,77	0	Computer Science	-	31%	33%
14	12,77	0	Computer Science	-	29%	33%
15	12,77	0	Medicine	-	27%	33%
16	12,81	0	Medicine	-	25%	33%
17	12,81	0	Architecture and Design	-	24%	33%
18	12,81	0	Computer Science	-	22%	33%
19	12,85	0	Computer Science	-	21%	33%
20	12,85	0	Medicine	-	20%	33%
21	12,85	0	Architecture and Design	-	19%	33%
22	12,85	0	Professional and Applied Computing	+	23%	42%
23	12,88	0	Medicine	-	22%	42%
24	12,88	0	Computer Science	-	21%	42%
25	12,88	0	Computer Science	-	20%	42%
26	12,88	0	Professional and Applied Computing	+	23%	50%
27	12,88	0	Computer Science	-	22%	50%
28	12,92	0	Professional and Applied Computing	+	25%	58%
29	12,92	0	Professional and Applied Computing	+	28%	67%
30	12,92	0	Professional and Applied Computing	+	30%	75%
31	12,92	0	Computer Science	-	29%	75%
32	12,92	0	Computer Science	-	28%	75%
33	12,92	0	Computer Science	-	27%	75%
34	12,92	0	Architecture and Design	-	26%	75%
35	12,92	0	Architecture and Design	-	26%	75%
36	12,92	0	Business and Economics	-	25%	75%
37	12,92	0	Business and Economics	-	24%	75%
38	12,92	0	Business and Economics	-	24%	75%
39	12,96	0	Computer Science	-	23%	75%
40	12,96	0	Computer Science	-	23%	75%
41	12,96	0	Business and Economics	-	22%	75%
42	12,96	0	Computer Science	-	21%	75%
43	12,96	0	Medicine	-	21%	75%
44	12,96	0	Business and Economics	-	20%	75%
45	12,96	0	Computer Science	-	20%	75%
46	12,96	0	Computer Science	-	20%	75%
47	13,00	0	Professional and Applied Computing	+	21%	83%
48	13,00	0	Computer Science	-	21%	83%
49	13,00	0	Professional and Applied Computing	+	22%	92%
50	13,00	0	Professional and Applied Computing	+	24%	100%

Foram encontradas 13 (treze) publicações sobre o assunto: "*Professional and Applied Computing*" e o gráfico da figura 4.10 apresenta a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve uma perda significativa, uma vez que inicialmente se tinha Precisão de 100% até o 3º documento e a partir da reclassificação por IAE, este índice se passou para 80% no 5º documento. A Recuperação que atingia o índice máximo no 50º documento se manteve. Houve perda de Precisão, que pode ser melhorado com uma melhor escolha das palavras relevantes ao assunto pesquisado.

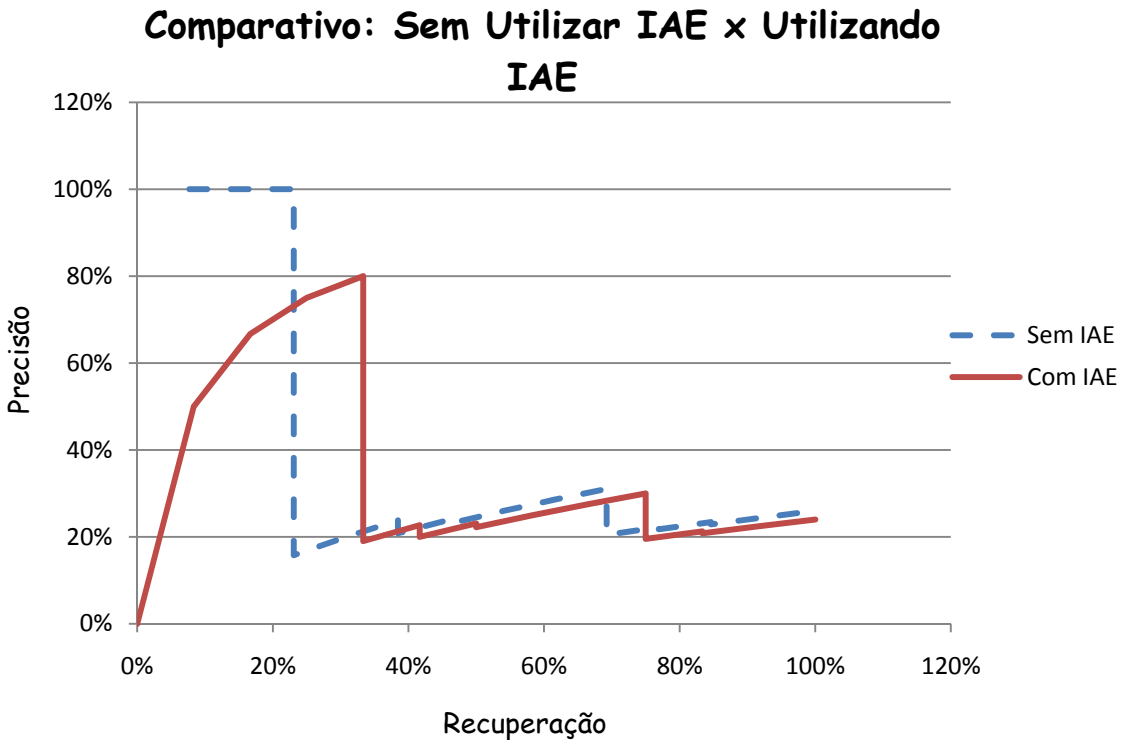


Figura 4.10 – Perfil 7: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma piora inicial na Precisão, mas que ocorreu uma melhora substancial na Recuperação de documentos. Concluí-se que a escolha dos termos mais relevantes para este assunto foi mal escolhido.

4.8 Perfil 8 – Assunto: *Medicine*

Para o oitavo perfil, foi escolhida uma publicação proveniente do *Springerlink*, inexistente na base de dados de testes, com assunto: "*Medicine*". A tabela 4.24 apresenta as características do perfil utilizado na pesquisa. As tabelas 4.25 e 4.26 apresentam os resultados obtidos pelo perfil a partir da classificação através do método *KNN* e a reclassificação através do Índice de Aproximação Estatística respectivamente. A relevância é positiva quando o assunto da publicação é "*Medicine*" e negativo para os demais casos nas tabelas 4.25 e 4.26.

Tabela 4.24 – Perfil 8 utilizado na pesquisa. Assunto: "*Medicine*".

Assunto: ***Medicine***	
Título: ***Epigastrica inferior Atypische Blutungsquelle bei Beckenfraktur***	
Palavras Relevantes:	**Peso**
1ª) *Medicine*	50%
2ª) *Diagnosis*	25%
3ª) *Pathology*	15%
4ª) *Anatomie*	10%
Resumo: The most common mechanism of pelvic fractures after blunt trauma is lateral commpression of the pelvis. Most of these fractures are of slight severity but it is possible, that a life-threatening hemorrhage can appear. The inferior epigastric artery is an atypical bleedingsite but it has to be considered in the search of the origin of the hemorrhage. Diagnostic tools are ultasound, computed tomography and angiography. After external fixation of the pelvis and persistent haemodynamic active bleeding is the percutaneous transcatheter embolization (PTE) in our opinion the treatment of choice. The introduced case is confirm with this statement.	

Tabela 4.25 – Perfil 8, classificação por *KNN*. Assunto: "*Medicine*".

Doc.	Distância	Asssunto	Relevância	Precisão	Recup.
1	9,00	Architecture and Design	-	0%	0%
2	9,00	Architecture and Design	-	0%	0%
3	9,00	Architecture and Design	-	0%	0%
4	9,00	Medicine	+	25%	4%
5	9,00	Medicine	+	40%	9%
6	9,00	Medicine	+	50%	13%
7	9,00	Biomedical and Life Sciences	-	43%	13%
8	9,06	Biomedical and Life Sciences	-	38%	13%
9	9,06	Architecture and Design	-	33%	13%
10	9,06	Biomedical and Life Sciences	-	30%	13%
11	9,06	Biomedical and Life Sciences	-	27%	13%
12	9,06	Biomedical and Life Sciences	-	25%	13%
13	9,06	Medicine	+	31%	17%
14	9,06	Medicine	+	36%	22%
15	9,06	Medicine	+	40%	26%
16	9,06	Medicine	+	44%	30%
17	9,06	Medicine	+	47%	35%
18	9,06	Medicine	+	50%	39%
19	9,06	Medicine	+	53%	43%
20	9,06	Medicine	+	55%	48%
21	9,06	Medicine	+	57%	52%
22	9,06	Computer Science	-	55%	52%
23	9,06	Medicine	+	57%	57%
24	9,06	Medicine	+	58%	61%
25	9,06	Computer Science	-	56%	61%
26	9,06	Computer Science	-	54%	61%
27	9,06	Professional and Applied Computing	-	52%	61%
28	9,06	Professional and Applied Computing	-	50%	61%
29	9,06	Professional and Applied Computing	-	48%	61%
30	9,06	Professional and Applied Computing	-	47%	61%
31	9,06	Professional and Applied Computing	-	45%	61%
32	9,06	Professional and Applied Computing	-	44%	61%
33	9,06	Professional and Applied Computing	-	42%	61%
34	9,06	Computer Science	-	41%	61%
35	9,06	Biomedical and Life Sciences	-	40%	61%
36	9,11	Biomedical and Life Sciences	-	39%	61%
37	9,11	Computer Science	-	38%	61%
38	9,11	Biomedical and Life Sciences	-	37%	61%
39	9,11	Biomedical and Life Sciences	-	36%	61%
40	9,11	Biomedical and Life Sciences	-	35%	61%
41	9,11	Biomedical and Life Sciences	-	34%	61%
42	9,11	Medicine	+	36%	65%
43	9,11	Medicine	+	37%	70%
44	9,11	Medicine	+	39%	74%
45	9,11	Medicine	+	40%	78%
46	9,11	Medicine	+	41%	83%
47	9,11	Medicine	+	43%	87%
48	9,11	Medicine	+	44%	91%
49	9,11	Medicine	+	45%	96%
50	9,11	Medicine	+	46%	100%

Tabela 4.26 – Perfil 8, reclassificação por IAE. Assunto: "*Medicine*".

Doc.	Distância	IAE	Assunto	Relevância	Precisão	Recup.
1	9,06	0,5	Medicine	+	100%	4%
2	9,11	0,5	Biomedical and Life Sciences	-	50%	4%
3	9,00	0	Architecture and Design	-	33%	4%
4	9,00	0	Architecture and Design	-	25%	4%
5	9,00	0	Architecture and Design	-	20%	4%
6	9,00	0	Medicine	+	33%	9%
7	9,00	0	Medicine	+	43%	13%
8	9,00	0	Medicine	+	50%	17%
9	9,00	0	Biomedical and Life Sciences	-	44%	17%
10	9,06	0	Biomedical and Life Sciences	-	40%	17%
11	9,06	0	Architecture and Design	-	36%	17%
12	9,06	0	Biomedical and Life Sciences	-	33%	17%
13	9,06	0	Biomedical and Life Sciences	-	31%	17%
14	9,06	0	Biomedical and Life Sciences	-	29%	17%
15	9,06	0	Medicine	+	33%	22%
16	9,06	0	Medicine	+	38%	26%
17	9,06	0	Medicine	+	41%	30%
18	9,06	0	Medicine	+	44%	35%
19	9,06	0	Medicine	+	47%	39%
20	9,06	0	Medicine	+	50%	43%
21	9,06	0	Medicine	+	52%	48%
22	9,06	0	Medicine	+	55%	52%
23	9,06	0	Computer Science	-	52%	52%
24	9,06	0	Medicine	+	54%	57%
25	9,06	0	Medicine	+	56%	61%
26	9,06	0	Computer Science	-	54%	61%
27	9,06	0	Computer Science	-	52%	61%
28	9,06	0	Professional and Applied Computing	-	50%	61%
29	9,06	0	Professional and Applied Computing	-	48%	61%
30	9,06	0	Professional and Applied Computing	-	47%	61%
31	9,06	0	Professional and Applied Computing	-	45%	61%
32	9,06	0	Professional and Applied Computing	-	44%	61%
33	9,06	0	Professional and Applied Computing	-	42%	61%
34	9,06	0	Professional and Applied Computing	-	41%	61%
35	9,06	0	Computer Science	-	40%	61%
36	9,06	0	Biomedical and Life Sciences	-	39%	61%
37	9,11	0	Biomedical and Life Sciences	-	38%	61%
38	9,11	0	Computer Science	-	37%	61%
39	9,11	0	Biomedical and Life Sciences	-	36%	61%
40	9,11	0	Biomedical and Life Sciences	-	35%	61%
41	9,11	0	Biomedical and Life Sciences	-	34%	61%
42	9,11	0	Medicine	+	36%	65%
43	9,11	0	Medicine	+	37%	70%
44	9,11	0	Medicine	+	39%	74%
45	9,11	0	Medicine	+	40%	78%
46	9,11	0	Medicine	+	41%	83%
47	9,11	0	Medicine	+	43%	87%
48	9,11	0	Medicine	+	44%	91%
49	9,11	0	Medicine	+	45%	96%
50	9,11	0	Medicine	+	46%	100%

Foram encontradas 23 (vinte e três) publicações sobre o assunto: "*Medicine*" e o gráfico da figura 4.11 apresenta a comparação entre a classificação por *KNN* e a reclassificação com IAE para as medidas de Precisão e Recuperação. Através dele verifica-se que a Precisão teve um ganho significativo, uma vez que inicialmente se tinha Precisão de 50% até o 6º documento e a partir da reclassificação por IAE, este índice chegou a 100% no 1º documento. A Recuperação que atingia o índice máximo no 50º documento se manteve. Houve ganho de Precisão que indica uma melhor escolha das palavras relevantes ao assunto pesquisado.

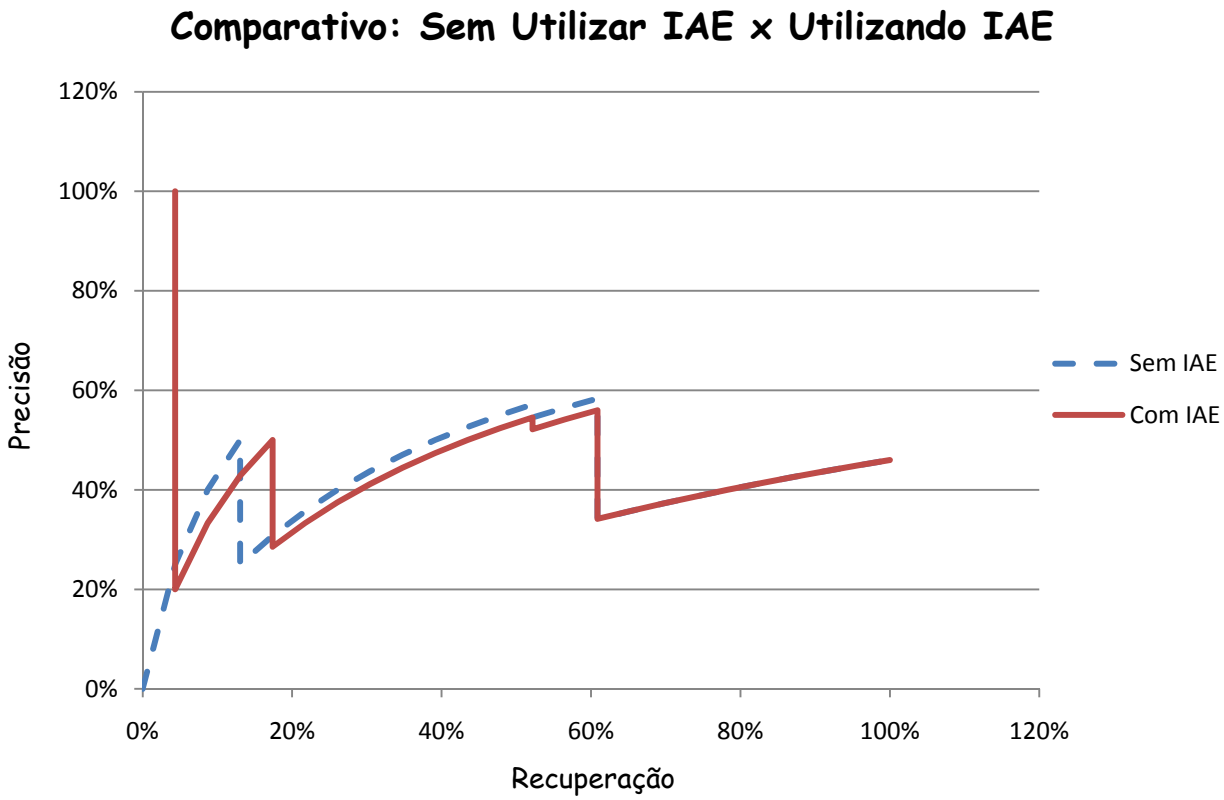


Figura 4.11 – Perfil 8: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) e da reclassificação (Com IAE).

Através do gráfico, nota-se que o uso do IAE produz sobre este perfil uma melhoria na Precisão e a Recuperação de documentos se manteve bem próximas em ambos os casos.

4.9 Conclusões e Problemas Encontrados no Uso da Metodologia

O uso do Índice de Aproximação Estatística melhorou os índices de Precisão e Recuperação, para os assuntos onde as quantidades de palavras relevantes na publicação eram maiores e os termos mais relevantes foram melhores escolhidos. Já nos casos onde as quantidades de palavras relevantes eram menores e os termos mais relevantes não foram bem escolhidos, não houve ganho substancial, inclusive ocorrendo um caso onde tanto a Precisão quanto a Recuperação de documentos ficaram com taxa 0 (zero).

4.10 Soluções Adotadas

Apesar da metodologia com o emprego do Índice de Aproximação Estatística se mostrar adequada quando do uso sobre assuntos com publicações de alta representatividade (alta média de palavras), acima de pelo menos 25 palavras, alguns ajustes se mostraram necessários para que este valor não fosse arbitrário.

Vários testes utilizando diferentes valores para a exclusão de publicações com número pequeno de palavras foram executados. Posteriormente testes com quantidades de palavras a taxa de 70% do valor médio também foram efetuados. Mesmo assim, ainda ocorreram casos em que não houve retorno de publicações pertinentes. Para estes testes foram necessários vários processamentos dos mesmos perfis com diferentes valores para a quantidade mínima de palavras, o que demandou muito tempo de processamento.

Na tabela 4.27 são apresentadas as médias de palavras das publicações por assunto.

Tabela 4.27 – Média de palavras por publicação.

Assunto	Total Palavras	Total Artigos	Média
Biomedical	1.022.921	19.753	51,8
Professional	268.001	8.646	30,9
Business	537.456	19.976	26,9
Architecture	120.765	2.873	42,0
Behavaioral	172.880	19.997	8,6
Computer	960.485	19.995	48,0
Medicine	174.749	13.323	13,1

A proposta adotada para a solução deste problema, que se mostrou eficaz e foi definida como Taxa de Similaridade (TS). A Taxa de Similaridade se constitui da contagem de palavras e suas relevâncias comuns entre o resumo do perfil usado e o resumo da publicação armazenada na base. A taxa é obtida pela divisão da contagem de palavras e suas respectivas ocorrências na publicação que também pertençam ao perfil, dividida pela quantidade de palavras e suas ocorrências no perfil. Assim, após o processamento pode-se efetuar pesquisas através de consultas ajustando a taxa de similaridade sem a necessidade de reprocessamentos.

A Taxa de Similaridade evita que publicações com poucas palavras relevantes sejam utilizadas, uma vez que, apesar das suas distâncias serem menores, não possuírem nenhuma similaridade com o perfil. O assunto *Behavarioral Science* por exemplo tem pouca representatividade e seus textos em grande maioria contém nomes, que não possuem representatividade para qualquer assunto. Apesar disso, se não levarmos em consideração a TS, as publicações referentes a este assunto tem distâncias muito pequenas para qualquer perfil que não seja muito extenso. A Taxa de Similaridade pode ser calculada através da equação:

$$\text{Taxa de Similaridade (TS)} = \frac{\sum (\text{word X ocorrência} \mid (\text{publicação}) \in \text{perfil})}{\sum (\text{word X ocorrência} \mid \text{perfil})} \qquad \text{(Eq. 9)}$$

Com esta proposta, deixaram de ser necessários novos processamentos para ajustes do número de palavras mínimas para que as publicações possam ser levadas em

conta na metodologia. A TS pode então ser usada para um ajuste durante a consulta sobre os resultados.

A seguir são apresentados os resultados sobre os perfis apresentados anteriormente, mas com o emprego da Taxa de Similaridade (TS) e a técnica de Palavras Freqüentes e Preditivas (PFP) proposta por POPESCUL e UNGAR [37]. Na tabela 4.28 os índices de Palavras Freqüentes e Preditivas calculados são apresentados e os 4 primeiros termos de maior valor foram extraídos para uso desta metodologia.

Com o uso do FPF, foi identificado o caso do termo "*die*", que apareceu em mais de um assunto. O mesmo foi desconsiderado por não dar precisão a um único assunto. Não é adequado para a metodologia que uma mesma palavra seja usada para identificar mais de um assunto, pois pode levar a ambigüidades.

Os perfis a seguir possuem os mesmos resumos dos perfis de 1 a 8, mas foram processados com o uso do número de palavras mínimas igual 8, com o emprego de PFP e passaram ser nomeados dos números 11 à 18 respectivamente. A todos eles foram aplicados nas consultas os coeficientes de Taxa de Similaridade de:

1. TS >= 0%;
2. TS > 0%;
3. TS >= 5%;
4. TS >= 10%;
5. TS >= 15%;
6. TS >= 20%.

Tabela 4.28 – Definição de termos por Palavras Freqüentes e Preditivas.

Assunto: *Architecture and Design*		
Termo	**PFP**	**Aproveitamento**
die	**2,92**	**0%**
sich	**2,65**	**25%**
architect	**2,61**	**24%**
den	**2,49**	**23%**
build	**2,41**	**23%**
Assunto: *Behavioral Science*		
Termo	**PFP**	**Aproveitamento**
book	**167,57**	**50%**
usawithout	**72,02**	**21%**
york	**57,70**	**17%**
new	**40,84**	**12%**
Assunto: *Biomedical and Life Sciences*		
Termo	**PFP**	**Aproveitamento**
gene	**246,13**	**39%**
sequenc	**165,04**	**26%**
marker	**115,79**	**18%**
chromosom	**111,76**	**17%**
Assunto: *Business and Economics*		
Termo	**PFP**	**Aproveitamento**
und	**109,52**	**46%**
ifr	**50,19**	**21%**
die	**48,77**	**0%**
angebot	**40,08**	**17%**
nachfrag	**40,05**	**17%**
Assunto: *Computer Science*		
Termo	**PFP**	**Aproveitamento**
algorithm	**110,50**	**28%**
paper	**98,34**	**25%**
system	**97,93**	**25%**
base	**92,65**	**23%**
Assunto: *Medicine*		
Termo	**PFP**	**Aproveitamento**
menschen	**22,02**	**30%**
der	**21,50**	**29%**
la	**15,74**	**21%**
sympathectomi	**15,09**	**20%**
Assunto: *Professional and Applied Computing*		
Termo	**PFP**	**Aproveitamento**
ll	**22,52**	**29%**
applic	**19,84**	**26%**
web	**18,27**	**24%**
server	**15,88**	**21%**

4.10.1 Perfil 11 – Assunto: *Computer Science*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.29 representa a definição do perfil 11 com PFP e TS.

Tabela 4.29 – Perfil 11 com PFP e TS.

Assunto: ***Computer Science***	
Título: ***Alpha Scale Spaces on a Bounded Domain***	
Palavras Relevantes:	**Peso**
1ª) *algorithm*	28%
2ª) *paper*	25%
3ª) *system*	25%
4ª) *base*	22%
Resumo: We consider a scale spaces, a parameterized class (a Î ([0,1]) of scale space representations beyond the well established Gaussian scale space, which are generated by the a th power of the minus Laplace operator on a bounded domain using the Neumann boundary condition. The Neumann boundary condition ensures that there is no grey value flux through the boundary. Thereby no artificial grey values from outside the image affect the evolution proces, which is the case for the a scale spaces on an unbounded domain. Moreover, the connection between the a scale spaces which is not trivial in the unbounded domain case, becomes straightforward The generator of the Gaussian semigroup extends to a compact, self adjoint operator on the Hilbert space \mathbb L 2(W) and therefore it has a complete countable set of eigen functions. Taking the a th power of the Gaussian generator simply boils down to taking the a th power of the corresponding eigenvalues. Consequently, all a scale spaces have exactly the same eigen modes and can be implemented simultaneously as scale dependent Fourier series. The only difference between them is the (relative) contribution of each eigen mode to the evolution proces. By introducing the notion of (non dimensional) relative scale in each a scale space, we are able to compare the various a scale spaces. The case a = 0.5, where the generator equals the square root of the minus Laplace operator leads to Poisson scale space, which is at least as interesting as Gaussian scale space and can be extended to a (Clifford) analytic scale space.	

As tabelas 4.30 e 4.31 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.12 e 4.13 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.30 – Comparativo das consultas do Perfil 11 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
13%	100%	4%	100%	3%	100%	2%	100%	2%	100%	13%	100%
13%	50%	4%	50%	6%	100%	5%	100%	5%	100%	25%	100%
13%	33%	9%	67%	9%	100%	7%	100%	7%	100%	38%	100%
13%	25%	9%	50%	9%	75%	7%	75%	10%	100%	50%	100%
25%	40%	9%	40%	12%	80%	9%	80%	12%	100%	63%	100%
25%	33%	13%	50%	15%	83%	12%	83%	15%	100%	63%	83%
38%	43%	13%	43%	18%	86%	14%	86%	17%	100%	63%	71%
38%	38%	17%	50%	21%	88%	16%	88%	20%	100%	75%	75%
50%	44%	22%	56%	24%	89%	19%	89%	22%	100%	75%	67%
63%	50%	22%	50%	24%	80%	19%	80%	24%	100%	75%	60%
63%	45%	22%	45%	24%	73%	21%	82%	24%	91%	75%	55%
63%	42%	22%	42%	24%	67%	23%	83%	27%	92%	75%	50%
63%	38%	26%	46%	24%	62%	26%	85%	29%	92%	88%	54%
63%	36%	26%	43%	24%	57%	28%	86%	29%	86%	88%	50%
63%	33%	26%	40%	24%	53%	30%	87%	32%	87%	100%	53%
63%	31%	30%	44%	26%	56%	33%	88%	34%	88%		
75%	35%	30%	41%	29%	59%	35%	88%	37%	88%		
75%	33%	35%	44%	32%	61%	37%	89%	39%	89%		
88%	37%	39%	47%	35%	63%	40%	89%	41%	89%		
88%	35%	43%	50%	38%	65%	42%	90%	41%	85%		
88%	33%	48%	52%	38%	62%	44%	90%	44%	86%		
100%	36%	48%	50%	41%	64%	47%	91%	46%	86%		
100%	35%	48%	48%	44%	65%	49%	91%	49%	87%		
100%	33%	48%	46%	47%	67%	51%	92%	51%	88%		
100%	32%	48%	44%	47%	64%	53%	92%	54%	88%		
100%	31%	52%	46%	47%	62%	56%	92%	56%	88%		
100%	30%	52%	44%	47%	59%	58%	93%	59%	89%		
100%	29%	52%	43%	50%	61%	58%	89%	61%	89%		
100%	28%	52%	41%	50%	59%	58%	86%	61%	86%		
100%	27%	52%	40%	50%	57%	60%	87%	63%	87%		
100%	26%	52%	39%	53%	58%	60%	84%	63%	84%		
100%	25%	52%	38%	56%	59%	63%	84%	66%	84%		
100%	24%	57%	39%	59%	61%	65%	85%	66%	82%		
100%	24%	61%	41%	62%	62%	67%	85%	66%	79%		
100%	23%	61%	40%	65%	63%	70%	86%	68%	80%		
100%	22%	65%	42%	68%	64%	70%	83%	71%	81%		
100%	22%	65%	41%	71%	65%	72%	84%	73%	81%		
100%	21%	70%	42%	74%	66%	72%	82%	76%	82%		
100%	21%	70%	41%	76%	67%	74%	82%	78%	82%		
100%	20%	74%	43%	79%	68%	77%	83%	80%	83%		
100%	20%	74%	41%	82%	68%	79%	83%	83%	83%		
100%	19%	78%	43%	85%	69%	81%	83%	85%	83%		
100%	19%	83%	44%	85%	67%	84%	84%	88%	84%		
100%	18%	87%	45%	85%	66%	86%	84%	88%	82%		
100%	18%	91%	47%	88%	67%	88%	84%	90%	82%		
100%	17%	96%	48%	88%	65%	91%	85%	93%	83%		
100%	17%	96%	47%	91%	66%	93%	85%	95%	83%		
100%	17%	96%	46%	94%	67%	95%	85%	98%	83%		
100%	16%	96%	45%	97%	67%	98%	86%	98%	82%		
100%	16%	100%	46%	100%	68%	100%	86%	100%	82%		

Tabela 4.31 – Comparativo das consultas do Perfil 11 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
13%	100%	4%	100%	3%	100%	2%	100%	2%	100%	13%	100%
13%	50%	9%	100%	6%	100%	5%	100%	5%	100%	13%	50%
13%	33%	9%	67%	9%	100%	7%	100%	5%	67%	13%	33%
25%	50%	9%	50%	12%	100%	9%	100%	7%	75%	25%	50%
25%	40%	13%	60%	12%	80%	12%	100%	10%	80%	38%	60%
25%	33%	17%	67%	12%	67%	14%	100%	12%	83%	50%	67%
25%	29%	17%	57%	15%	71%	16%	100%	15%	86%	50%	57%
38%	38%	22%	63%	18%	75%	19%	100%	17%	88%	50%	50%
38%	33%	22%	56%	21%	78%	21%	100%	20%	89%	63%	56%
50%	40%	26%	60%	24%	80%	23%	100%	22%	90%	75%	60%
50%	36%	26%	55%	26%	82%	26%	100%	22%	82%	75%	55%
63%	42%	30%	58%	29%	83%	28%	100%	24%	83%	88%	58%
63%	38%	35%	62%	32%	85%	30%	100%	27%	85%	100%	62%
63%	36%	39%	64%	35%	86%	33%	100%	29%	86%	100%	57%
63%	33%	43%	67%	38%	87%	35%	100%	32%	87%	100%	53%
63%	31%	48%	69%	41%	88%	37%	100%	34%	88%		
63%	29%	52%	71%	44%	88%	40%	100%	37%	88%		
75%	33%	57%	72%	47%	89%	42%	100%	37%	83%		
88%	37%	57%	68%	50%	89%	42%	95%	39%	84%		
88%	35%	61%	70%	53%	90%	44%	95%	41%	85%		
88%	33%	65%	71%	56%	90%	47%	95%	44%	86%		
100%	36%	65%	68%	59%	91%	49%	95%	46%	86%		
100%	35%	70%	70%	59%	87%	51%	96%	49%	87%		
100%	33%	70%	67%	59%	83%	53%	96%	51%	88%		
100%	32%	70%	64%	62%	84%	56%	96%	54%	88%		
100%	31%	74%	65%	65%	85%	58%	96%	56%	88%		
100%	30%	78%	67%	65%	81%	60%	96%	59%	89%		
100%	29%	78%	64%	68%	82%	63%	96%	61%	89%		
100%	28%	78%	62%	68%	79%	65%	97%	63%	90%		
100%	27%	78%	60%	68%	77%	67%	97%	66%	90%		
100%	26%	83%	61%	71%	77%	70%	97%	68%	90%		
100%	25%	87%	63%	74%	78%	72%	97%	71%	91%		
100%	24%	87%	61%	74%	76%	74%	97%	71%	88%		
100%	24%	87%	59%	74%	74%	77%	97%	71%	85%		
100%	23%	87%	57%	76%	74%	79%	97%	73%	86%		
100%	22%	87%	56%	79%	75%	81%	97%	73%	83%		
100%	22%	91%	57%	82%	76%	84%	97%	73%	81%		
100%	21%	91%	55%	82%	74%	86%	97%	76%	82%		
100%	21%	91%	54%	82%	72%	88%	97%	78%	82%		
100%	20%	91%	53%	85%	73%	88%	95%	80%	83%		
100%	20%	91%	51%	85%	71%	91%	95%	83%	83%		
100%	19%	91%	50%	85%	69%	91%	93%	85%	83%		
100%	19%	91%	49%	85%	67%	93%	93%	88%	84%		
100%	18%	96%	50%	88%	68%	93%	91%	90%	84%		
100%	18%	96%	49%	91%	69%	93%	89%	93%	84%		
100%	17%	96%	48%	94%	70%	95%	89%	95%	85%		
100%	17%	96%	47%	97%	70%	95%	87%	95%	83%		
100%	17%	100%	48%	100%	71%	95%	85%	98%	83%		
100%	16%	100%	47%	100%	69%	98%	86%	100%	84%		
100%	16%	100%	46%	100%	68%	100%	86%	100%	82%		

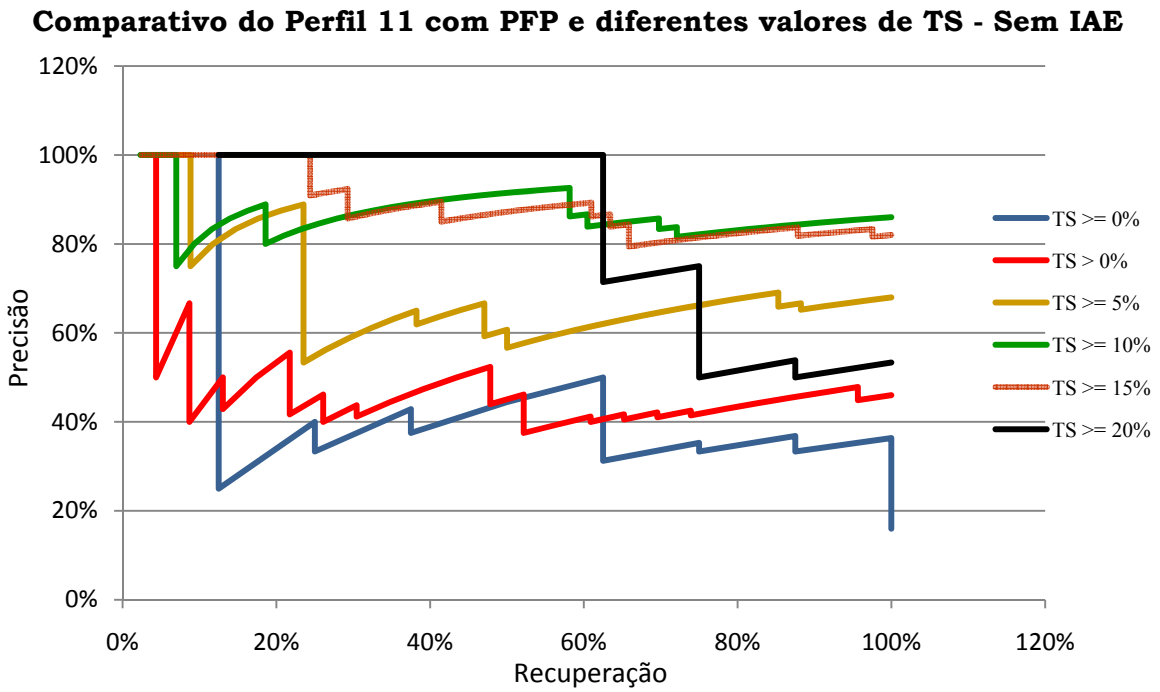


Figura 4.12 – Perfil 11: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

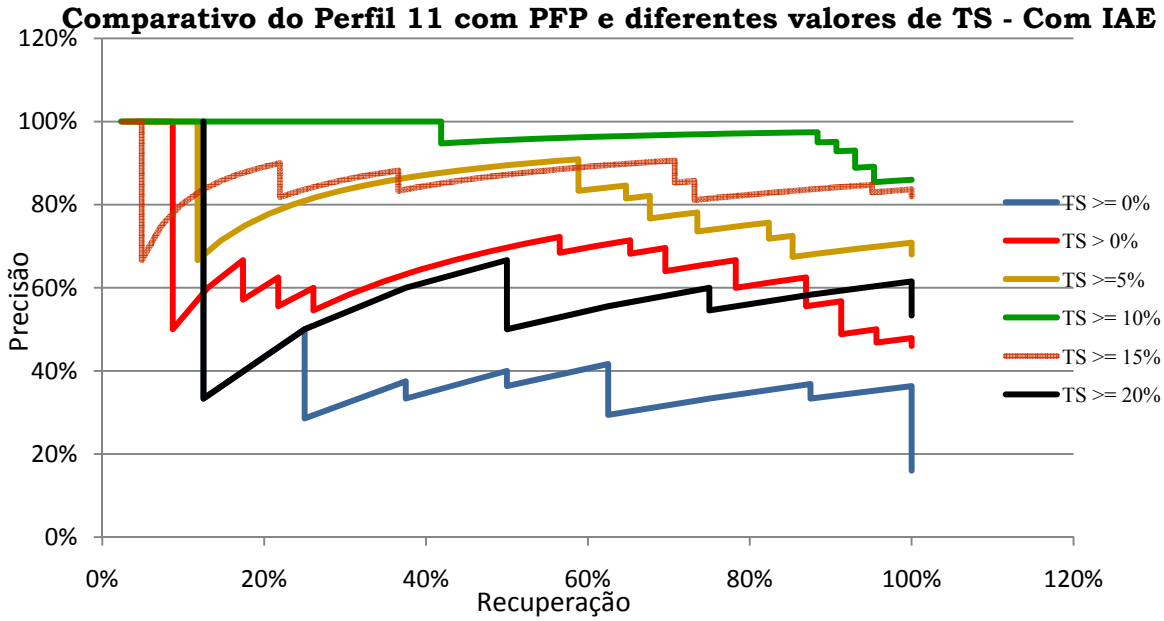


Figura 4.13 – Perfil 11: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Note que sem o uso do IAE, os melhores resultados foram obtidos sobre as consultas com 20% ou mais de Taxa de Similaridade, mas com o emprego do IAE os melhores resultados foram obtidos utilizando uma Taxa de Similaridade bem menor, de 10% ou mais. Ainda permitiu que um número maior de documentos fosse recuperado sobre o assunto definido.

4.10.2 Perfil 12 – Assunto: *Computer Science*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.32 representa a definição do perfil 12 com PFP e TS.

Tabela 4.32 – Perfil 12 com PFP e TS.

Assunto: ***Computer Science***	
Título: ***An interface based on transputers to simulate the dynamic equation of robot manipulators using parallel computing***	
Palavras Relevantes:	**Peso**
1ª) *algorithm*	28%
2ª) *paper*	25%
3ª) *system*	25%
4ª) *base*	22%
Resumo: The inverse dynamics control of robot manipulators is based on the application of a nonlinear feedback control law. The implementation of this control requires the computation of all the terms of the dynamic equation at each sample instant. In order to solve this problem, distributed memory parallel algorithms using the Lagrange-Euler formulation are presented. This formulation permits us to establish matrix structures that are distributed among the processors by rows with a good computational load balance. A new Windows interface, called WinServer, based on transputers, which permits us to modify the robot parameters and to simulate the dynamic equation is also presented. This interface can be used as a monitoring tool that allows the user to know the situation of the different processes.	

As tabelas 4.33 e 4.34 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.14 e 4.15 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.33 – Comparativo das consultas do Perfil 12 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	3%	100%	3%	100%	3%	100%	2%	100%
0%	0%	0%	0%	5%	100%	3%	50%	5%	100%	5%	100%
0%	0%	0%	0%	8%	100%	5%	67%	8%	100%	7%	100%
0%	0%	4%	25%	11%	100%	8%	75%	11%	100%	9%	100%
0%	0%	7%	40%	14%	100%	10%	80%	13%	100%	11%	100%
0%	0%	7%	33%	16%	100%	13%	83%	16%	100%	14%	100%
6%	14%	7%	29%	19%	100%	15%	86%	18%	100%	16%	100%
11%	25%	7%	25%	19%	88%	18%	88%	21%	100%	18%	100%
11%	22%	11%	33%	19%	78%	21%	89%	24%	100%	20%	100%
11%	20%	14%	40%	22%	80%	23%	90%	24%	90%	23%	100%
11%	18%	18%	45%	22%	73%	26%	91%	26%	91%	25%	100%
11%	17%	21%	50%	22%	67%	28%	92%	29%	92%	27%	100%
11%	15%	21%	46%	22%	62%	28%	85%	32%	92%	30%	100%
17%	21%	25%	50%	24%	64%	31%	86%	34%	93%	32%	100%
22%	27%	25%	47%	24%	60%	31%	80%	37%	93%	34%	100%
28%	31%	29%	50%	24%	56%	33%	81%	37%	88%	36%	100%
33%	35%	29%	47%	27%	59%	36%	82%	37%	82%	39%	100%
33%	33%	29%	44%	30%	61%	38%	83%	39%	83%	41%	100%
33%	32%	32%	47%	30%	58%	41%	84%	39%	79%	43%	100%
33%	30%	36%	50%	32%	60%	44%	85%	39%	75%	45%	100%
33%	29%	39%	52%	35%	62%	46%	86%	39%	71%	45%	95%
33%	27%	43%	55%	38%	64%	49%	86%	39%	68%	48%	95%
33%	26%	46%	57%	41%	65%	51%	87%	39%	65%	50%	96%
33%	25%	50%	58%	43%	67%	54%	88%	42%	67%	52%	96%
33%	24%	54%	60%	43%	64%	54%	84%	45%	68%	55%	96%
33%	23%	57%	62%	46%	65%	56%	85%	47%	69%	57%	96%
33%	22%	61%	63%	49%	67%	59%	85%	50%	70%	59%	96%
33%	21%	61%	61%	51%	68%	62%	86%	53%	71%	61%	96%
33%	21%	61%	59%	54%	69%	64%	86%	55%	72%	64%	97%
33%	20%	61%	57%	57%	70%	67%	87%	58%	73%	66%	97%
39%	23%	64%	58%	59%	71%	69%	87%	61%	74%	68%	97%
39%	22%	64%	56%	62%	72%	72%	88%	61%	72%	70%	97%
44%	24%	64%	55%	65%	73%	74%	88%	61%	70%	73%	97%
44%	24%	64%	53%	68%	74%	77%	88%	63%	71%	75%	97%
44%	23%	68%	54%	70%	74%	77%	86%	66%	71%	75%	94%
44%	22%	71%	56%	70%	72%	79%	86%	66%	69%	77%	94%
50%	24%	75%	57%	70%	70%	82%	86%	68%	70%	80%	95%
56%	26%	75%	55%	70%	68%	82%	84%	71%	71%	80%	92%
61%	28%	75%	54%	73%	69%	82%	82%	74%	72%	80%	90%
67%	30%	75%	53%	76%	70%	82%	80%	76%	73%	82%	90%
72%	32%	79%	54%	78%	71%	85%	80%	79%	73%	84%	90%
78%	33%	82%	55%	81%	71%	87%	81%	82%	74%	86%	90%
83%	35%	86%	56%	84%	72%	90%	81%	84%	74%	86%	88%
83%	34%	86%	55%	86%	73%	92%	82%	84%	73%	89%	89%
89%	36%	86%	53%	89%	73%	92%	80%	87%	73%	91%	89%
94%	37%	89%	54%	92%	74%	95%	80%	89%	74%	91%	87%
94%	36%	93%	55%	95%	74%	97%	81%	92%	74%	93%	87%
94%	35%	96%	56%	97%	75%	97%	79%	95%	75%	95%	88%
94%	35%	100%	57%	97%	73%	97%	78%	97%	76%	98%	88%
100%	36%	100%	56%	100%	74%	100%	78%	100%	76%	100%	88%

Tabela 4.34 – Comparativo das consultas do Perfil 12 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
6%	100%	4%	100%	3%	100%	3%	100%	3%	100%	2%	100%
11%	100%	7%	100%	3%	50%	5%	100%	5%	100%	5%	100%
17%	100%	11%	100%	5%	67%	8%	100%	8%	100%	7%	100%
22%	100%	14%	100%	8%	75%	10%	100%	11%	100%	9%	100%
28%	100%	18%	100%	11%	80%	13%	100%	13%	100%	11%	100%
33%	100%	21%	100%	14%	83%	15%	100%	16%	100%	14%	100%
39%	100%	25%	100%	16%	86%	15%	86%	18%	100%	14%	86%
44%	100%	29%	100%	19%	88%	18%	88%	21%	100%	16%	88%
44%	89%	32%	100%	22%	89%	21%	89%	24%	100%	18%	89%
44%	80%	36%	100%	24%	90%	23%	90%	26%	100%	20%	90%
44%	73%	39%	100%	27%	91%	26%	91%	29%	100%	20%	82%
44%	67%	39%	92%	30%	92%	28%	92%	32%	100%	23%	83%
44%	62%	43%	92%	32%	92%	31%	92%	34%	100%	25%	85%
44%	57%	43%	86%	32%	86%	33%	93%	37%	100%	27%	86%
44%	53%	43%	80%	35%	87%	36%	93%	39%	100%	30%	87%
50%	56%	43%	75%	38%	88%	38%	94%	42%	100%	32%	88%
50%	53%	43%	71%	41%	88%	41%	94%	45%	100%	34%	88%
50%	50%	46%	72%	43%	89%	44%	94%	47%	100%	36%	89%
50%	47%	50%	74%	46%	89%	46%	95%	50%	100%	39%	89%
50%	45%	54%	75%	49%	90%	49%	95%	53%	100%	41%	90%
50%	43%	54%	71%	51%	90%	51%	95%	55%	100%	43%	90%
56%	45%	54%	68%	54%	91%	54%	95%	58%	100%	45%	91%
61%	48%	57%	70%	57%	91%	56%	96%	61%	100%	48%	91%
67%	50%	57%	67%	59%	92%	59%	96%	63%	100%	50%	92%
67%	48%	57%	64%	62%	92%	62%	96%	66%	100%	52%	92%
67%	46%	57%	62%	65%	92%	64%	96%	68%	100%	55%	92%
67%	44%	61%	63%	68%	93%	67%	96%	71%	100%	57%	93%
67%	43%	64%	64%	70%	93%	69%	96%	74%	100%	59%	93%
67%	41%	68%	66%	73%	93%	72%	97%	76%	100%	61%	93%
67%	40%	68%	63%	76%	93%	74%	97%	79%	100%	64%	93%
67%	39%	68%	61%	78%	94%	77%	97%	82%	100%	66%	94%
67%	38%	68%	59%	81%	94%	79%	97%	84%	100%	68%	94%
67%	36%	71%	61%	84%	94%	82%	97%	87%	100%	70%	94%
67%	35%	75%	62%	84%	91%	85%	97%	89%	100%	73%	94%
67%	34%	79%	63%	84%	89%	87%	97%	89%	97%	75%	94%
67%	33%	82%	64%	84%	86%	90%	97%	92%	97%	77%	94%
67%	32%	82%	62%	86%	86%	92%	97%	95%	97%	80%	95%
67%	32%	82%	61%	89%	87%	92%	95%	97%	97%	82%	95%
67%	31%	82%	59%	89%	85%	95%	95%	97%	95%	84%	95%
67%	30%	82%	58%	89%	83%	97%	95%	97%	93%	86%	95%
72%	32%	86%	59%	89%	80%	100%	95%	97%	90%	89%	95%
78%	33%	86%	57%	92%	81%	100%	93%	97%	88%	91%	95%
83%	35%	86%	56%	92%	79%	100%	91%	97%	86%	93%	95%
89%	36%	89%	57%	95%	80%	100%	89%	97%	84%	95%	95%
94%	38%	89%	56%	97%	80%	100%	87%	97%	82%	95%	93%
94%	37%	89%	54%	97%	78%	100%	85%	100%	83%	98%	93%
100%	38%	93%	55%	97%	77%	100%	83%	100%	81%	98%	91%
100%	38%	96%	56%	97%	75%	100%	81%	100%	79%	98%	90%
100%	37%	100%	57%	97%	73%	100%	80%	100%	78%	98%	88%
100%	36%	100%	56%	100%	74%	100%	78%	100%	76%	100%	88%

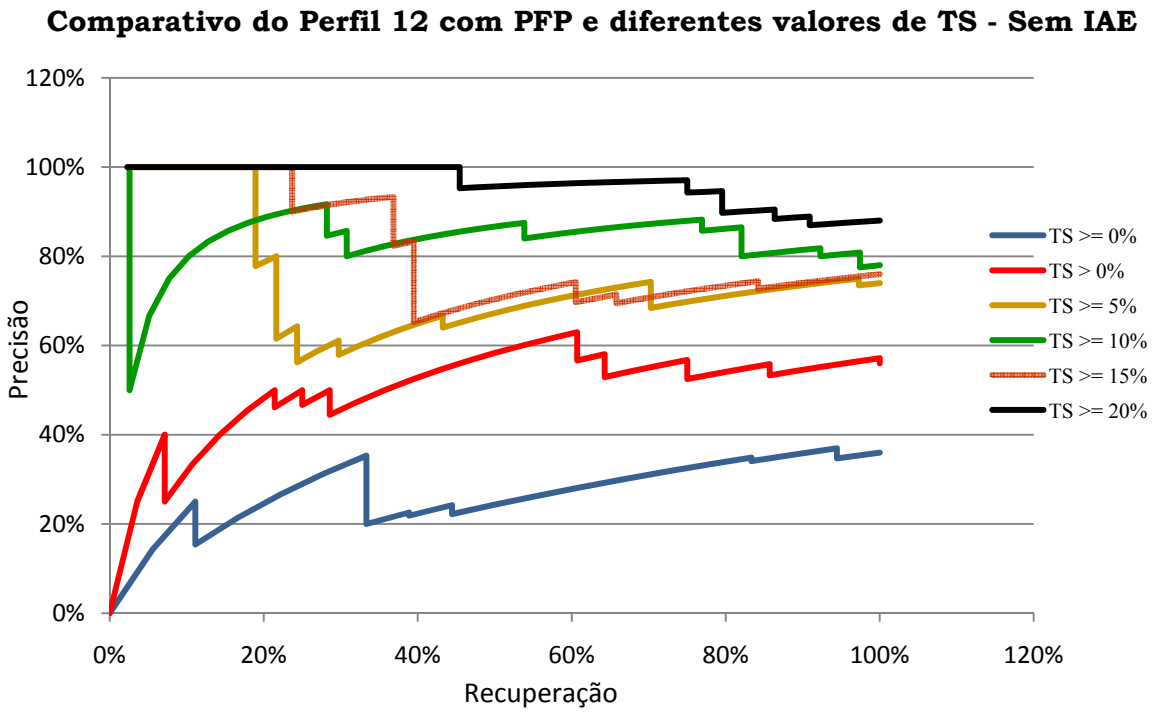


Figura 4.14 – Perfil 12: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

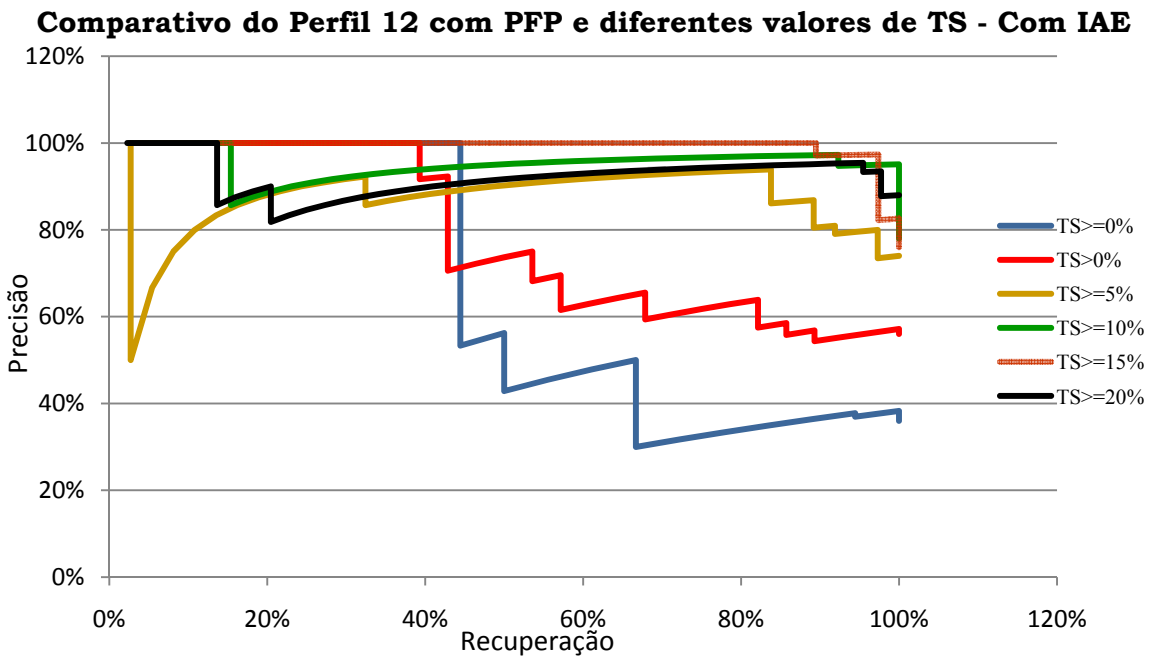


Figura 4.15 – Perfil 12: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Note que sem o uso do IAE, os melhores resultados foram obtidos sobre as consultas com 20% ou mais de Taxa de Similaridade, mas com o emprego do IAE os melhores resultados foram obtidos utilizando uma Taxa de Similaridade bem menor, de 15% ou mais. Mesmo assim, o emprego de Taxas de Similaridade de 10% e 5% ou mais e acima de 0% tiveram excelente desempenho.

Perfil 13 – Assunto: *Biomedical and Life Sciences*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.35 representa a definição do perfil 13 com PFP e TS.

Tabela 4.35 – Perfil 13 com PFP e TS.

Assunto: ***Biomedical and Life Science***	
Título: ***A geography of ecosystem vulnerability***	
Palavras Relevantes:	**Peso**
1ª) *gene*	39%
2ª) *sequenc*	26%
3ª) *marker*	18%
4ª) *chromosom*	17%
Resumo: Land-cover change and the subsequent potential loss of natural resources due to conversion to anthropogenic use is regarded as one of the more pervasive environmental threats. Population and road data were used to generate interpolated surfaces of land demand across a large region, the mid-Atlantic states of Pennsylvania, Delaware, Maryland, Virginia, and West Virginia. The land demand surfaces were evaluated against land-cover change, as estimated using temporal decline in Normalized Difference Vegetation Index (NDVI). In general, the interpolated surfaces exhibited a plateau along the eastern seaboard that sank to a valley in the center of the study area, and then rose again to a plateau in the west that was of overall lower height than the plateau on the eastern seaboard. The spatial pattern of land-cover change showed the same general pattern as the interpolated surfaces of land demand. Correlations were significant regardless of variations used to generate the interpolated surfaces. The results suggest that human activity is the principal agent of land-cover change at regional scales in this region, and that natural resources that change as land cover changes (e.g., water, habitat) are exposed to a gradient of vulnerability that increases from west to east.	

As tabelas 4.36 e 4.37 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.16 e 4.17 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.36 – Comparativo das consultas do Perfil 13 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	100%
0%	0%	0%	0%	0%	0%	0%	0%	7%	50%	3%	50%
14%	33%	0%	0%	0%	0%	0%	0%	7%	33%	3%	33%
14%	25%	0%	0%	0%	0%	0%	0%	7%	25%	3%	25%
14%	20%	0%	0%	0%	0%	0%	0%	7%	20%	6%	40%
14%	17%	0%	0%	0%	0%	0%	0%	7%	17%	10%	50%
14%	14%	0%	0%	0%	0%	0%	0%	7%	14%	13%	57%
14%	13%	0%	0%	0%	0%	0%	0%	7%	13%	16%	63%
14%	11%	0%	0%	0%	0%	0%	0%	7%	11%	16%	56%
14%	10%	0%	0%	0%	0%	0%	0%	7%	10%	19%	60%
14%	9%	0%	0%	0%	0%	0%	0%	7%	9%	19%	55%
14%	8%	0%	0%	0%	0%	0%	0%	7%	8%	19%	50%
14%	8%	14%	8%	50%	8%	0%	0%	7%	8%	23%	54%
29%	14%	29%	14%	50%	7%	0%	0%	7%	7%	26%	57%
43%	20%	29%	13%	50%	7%	0%	0%	7%	7%	29%	60%
57%	25%	29%	13%	50%	6%	0%	0%	14%	13%	29%	56%
57%	24%	29%	12%	50%	6%	0%	0%	21%	18%	32%	59%
57%	22%	29%	11%	50%	6%	0%	0%	21%	17%	35%	61%
57%	21%	29%	11%	50%	5%	0%	0%	21%	16%	39%	63%
57%	20%	29%	10%	50%	5%	0%	0%	21%	15%	42%	65%
57%	19%	29%	10%	50%	5%	0%	0%	29%	19%	45%	67%
57%	18%	29%	9%	50%	5%	0%	0%	29%	18%	45%	64%
57%	17%	29%	9%	50%	4%	0%	0%	29%	17%	48%	65%
57%	17%	43%	13%	50%	4%	0%	0%	29%	17%	48%	63%
71%	20%	43%	12%	50%	4%	0%	0%	29%	16%	52%	64%
86%	23%	43%	12%	50%	4%	0%	0%	29%	15%	55%	65%
86%	22%	43%	11%	50%	4%	0%	0%	29%	15%	55%	63%
100%	25%	43%	11%	50%	4%	0%	0%	29%	14%	58%	64%
100%	24%	43%	10%	50%	3%	0%	0%	29%	14%	61%	66%
100%	23%	43%	10%	50%	3%	0%	0%	29%	13%	65%	67%
100%	23%	57%	13%	50%	3%	0%	0%	36%	16%	68%	68%
100%	22%	71%	16%	50%	3%	0%	0%	36%	16%	68%	66%
100%	21%	71%	15%	50%	3%	0%	0%	36%	15%	68%	64%
100%	21%	71%	15%	50%	3%	0%	0%	36%	15%	68%	62%
100%	20%	71%	14%	50%	3%	0%	0%	43%	17%	68%	60%
100%	19%	71%	14%	100%	6%	0%	0%	50%	19%	71%	61%
100%	19%	71%	14%	100%	5%	0%	0%	57%	22%	74%	62%
100%	18%	71%	13%	100%	5%	33%	3%	57%	21%	77%	63%
100%	18%	71%	13%	100%	5%	33%	3%	57%	21%	77%	62%
100%	18%	71%	13%	100%	5%	33%	3%	64%	23%	77%	60%
100%	17%	86%	15%	100%	5%	33%	2%	64%	22%	77%	59%
100%	17%	86%	14%	100%	5%	33%	2%	64%	21%	81%	60%
100%	16%	86%	14%	100%	5%	33%	2%	71%	23%	81%	58%
100%	16%	86%	14%	100%	5%	33%	2%	79%	25%	81%	57%
100%	16%	86%	13%	100%	4%	33%	2%	86%	27%	84%	58%
100%	15%	100%	15%	100%	4%	33%	2%	93%	28%	87%	59%
100%	15%	100%	15%	100%	4%	67%	4%	100%	30%	90%	60%
100%	15%	100%	15%	100%	4%	100%	6%	100%	29%	94%	60%
100%	14%	100%	14%	100%	4%	100%	6%	100%	29%	97%	61%
100%	14%	100%	14%	100%	4%	100%	6%	100%	28%	100%	62%

Tabela 4.37 – Comparativo das consultas do Perfil 13 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	7%	100%	3%	100%
0%	0%	0%	0%	0%	0%	33%	50%	7%	50%	6%	100%
14%	33%	0%	0%	0%	0%	33%	33%	7%	33%	10%	100%
14%	25%	0%	0%	0%	0%	33%	25%	7%	25%	13%	100%
14%	20%	0%	0%	0%	0%	33%	20%	7%	20%	16%	100%
14%	17%	0%	0%	0%	0%	33%	17%	14%	33%	16%	83%
14%	14%	0%	0%	0%	0%	33%	14%	14%	29%	16%	71%
14%	13%	0%	0%	0%	0%	33%	13%	21%	38%	16%	63%
14%	11%	0%	0%	0%	0%	33%	11%	21%	33%	19%	67%
14%	10%	0%	0%	0%	0%	33%	10%	21%	30%	23%	70%
14%	9%	0%	0%	0%	0%	33%	9%	21%	27%	26%	73%
14%	8%	0%	0%	0%	0%	33%	8%	21%	25%	29%	75%
14%	8%	14%	8%	0%	0%	33%	8%	21%	23%	29%	69%
29%	14%	29%	14%	50%	7%	33%	7%	21%	21%	32%	71%
43%	20%	29%	13%	50%	7%	33%	7%	21%	20%	32%	67%
57%	25%	29%	13%	50%	6%	33%	6%	21%	19%	32%	63%
57%	24%	29%	12%	50%	6%	33%	6%	21%	18%	35%	65%
57%	22%	29%	11%	50%	6%	33%	6%	21%	17%	39%	67%
57%	21%	29%	11%	50%	5%	33%	5%	21%	16%	42%	68%
57%	20%	29%	10%	50%	5%	33%	5%	29%	20%	42%	65%
57%	19%	29%	10%	50%	5%	33%	5%	29%	19%	45%	67%
57%	18%	29%	9%	50%	5%	33%	5%	29%	18%	48%	68%
57%	17%	29%	9%	50%	4%	33%	4%	29%	17%	52%	70%
57%	17%	43%	13%	50%	4%	33%	4%	36%	21%	55%	71%
71%	20%	43%	12%	50%	4%	33%	4%	36%	20%	58%	72%
86%	23%	43%	12%	50%	4%	33%	4%	36%	19%	58%	69%
86%	22%	43%	11%	50%	4%	33%	4%	36%	19%	61%	70%
100%	25%	43%	11%	50%	4%	33%	4%	36%	18%	61%	68%
100%	24%	43%	10%	50%	3%	33%	3%	36%	17%	65%	69%
100%	23%	43%	10%	50%	3%	33%	3%	36%	17%	65%	67%
100%	23%	57%	13%	50%	3%	33%	3%	36%	16%	68%	68%
100%	22%	71%	16%	50%	3%	33%	3%	36%	16%	71%	69%
100%	21%	71%	15%	50%	3%	33%	3%	43%	18%	71%	67%
100%	21%	71%	15%	50%	3%	33%	3%	43%	18%	71%	65%
100%	20%	71%	14%	50%	3%	33%	3%	43%	17%	71%	63%
100%	19%	71%	14%	50%	3%	33%	3%	50%	19%	71%	61%
100%	19%	71%	14%	100%	5%	33%	3%	57%	22%	74%	62%
100%	18%	71%	13%	100%	5%	33%	3%	64%	24%	77%	63%
100%	18%	71%	13%	100%	5%	67%	5%	64%	23%	77%	62%
100%	18%	71%	13%	100%	5%	67%	5%	64%	23%	77%	60%
100%	17%	86%	15%	100%	5%	67%	5%	71%	24%	77%	59%
100%	17%	86%	14%	100%	5%	67%	5%	71%	24%	81%	60%
100%	16%	86%	14%	100%	5%	67%	5%	71%	23%	81%	58%
100%	16%	86%	14%	100%	5%	67%	5%	79%	25%	81%	57%
100%	16%	86%	13%	100%	4%	67%	4%	86%	27%	84%	58%
100%	15%	100%	15%	100%	4%	67%	4%	93%	28%	87%	59%
100%	15%	100%	15%	100%	4%	67%	4%	100%	30%	90%	60%
100%	15%	100%	15%	100%	4%	100%	6%	100%	29%	94%	60%
100%	14%	100%	14%	100%	4%	100%	6%	100%	29%	97%	61%
100%	14%	100%	14%	100%	4%	100%	6%	100%	28%	100%	62%

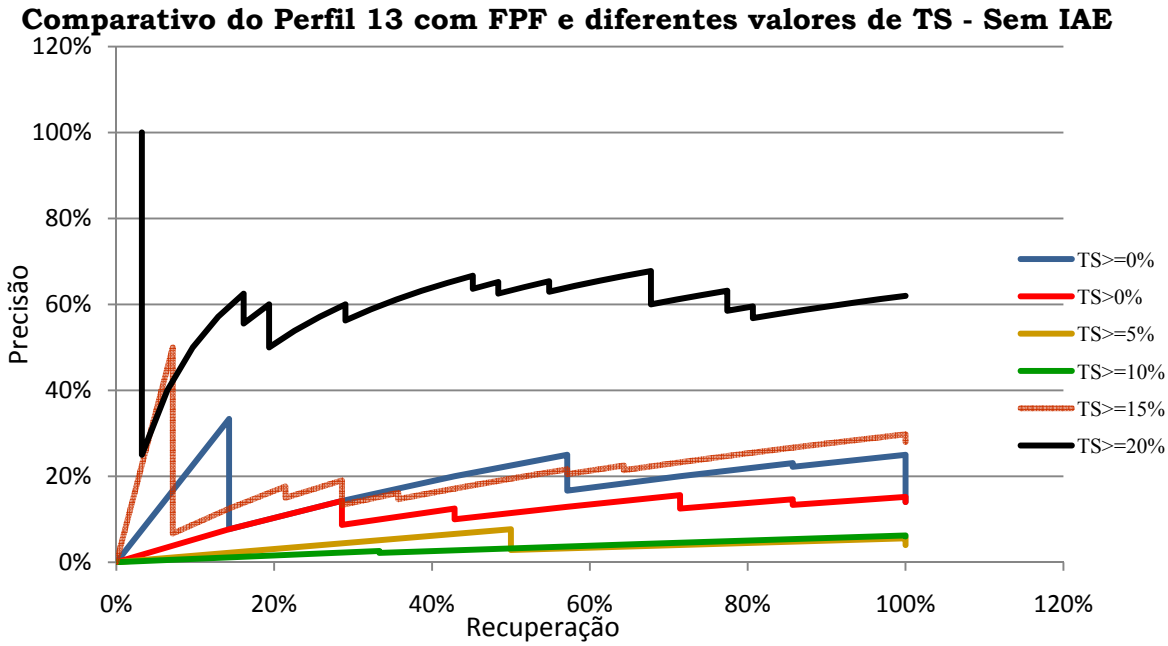


Figura 4.16 – Perfil 13: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

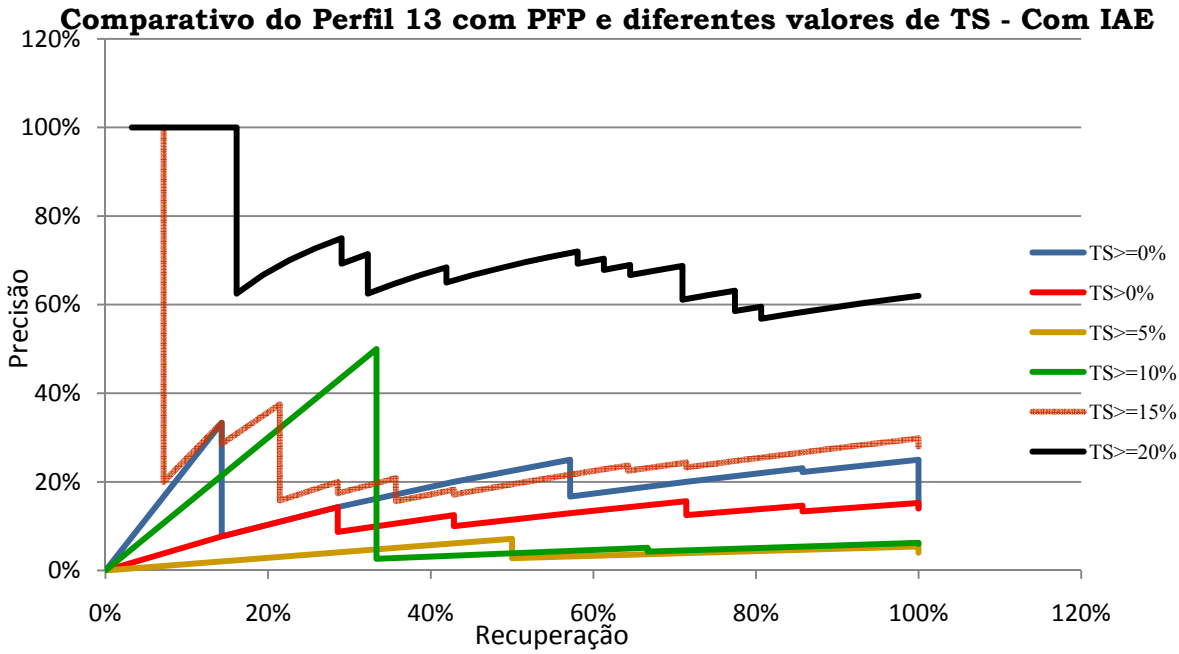


Figura 4.17 – Perfil 13: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Note que sem o uso do IAE, os melhores resultados foram obtidos sobre as consultas com 20% ou mais de Taxa de Similaridade, mas com o emprego do IAE os todos os resultados tiveram melhorias, com exceção a consulta com Taxa de Similaridade de 5% ou mais que se manteve estável.

Perfil 14 – Assunto: *Behavioral Science*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.38 representa a definição do perfil 14 com PFP e TS.

Tabela 4.38 – Perfil 14 com PFP e TS.

Assunto: ***Behavioral Science***	
Título: ***A study in word formation restrictions***	
Palavras Relevantes:	**Peso**
1ª) *book*	50%
2ª) *without*	21%
3ª) *york*	17%
4ª) *new*	12%
Resumo: Subjects judged the acceptability of pseudo words formed a stem, an optional affix, and-ity. The results show that those words in which-ity was adjacent to a latinate morpheme were more acceptable than those in which it was not. Other factors which influenced the acceptability ratings were familiarity and homogeneity of morpheme types. In their evaluations of the possible words, subjects were sensitive to historical class distinctions of stems and affixes.	

As tabelas 4.39 e 4.40 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.18 e 4.19 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.39 – Comparativo das consultas do Perfil 14 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
11%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
14%	29%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17%	33%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
19%	37%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
22%	40%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
25%	43%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
28%	45%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
31%	48%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36%	52%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36%	48%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
39%	50%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
42%	52%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
44%	53%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
47%	55%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
50%	56%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
53%	58%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
56%	59%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
58%	60%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
61%	61%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
64%	62%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
67%	63%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
69%	64%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
72%	65%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
75%	66%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
78%	67%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
81%	67%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
83%	68%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
86%	69%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
89%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
92%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
94%	71%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
97%	71%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	72%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Tabela 4.40 – Comparativo das consultas do Perfil 14 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
3%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
8%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
11%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
14%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
19%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
22%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
25%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
28%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
31%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
36%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
39%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
42%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
44%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
47%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
50%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
53%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
56%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
58%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
61%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
64%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
67%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
69%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
72%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
75%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
78%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
81%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
83%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
86%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
89%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
92%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
94%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
97%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	97%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	95%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	88%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	86%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	84%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	82%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	80%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	78%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	77%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	75%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	73%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
100%	72%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

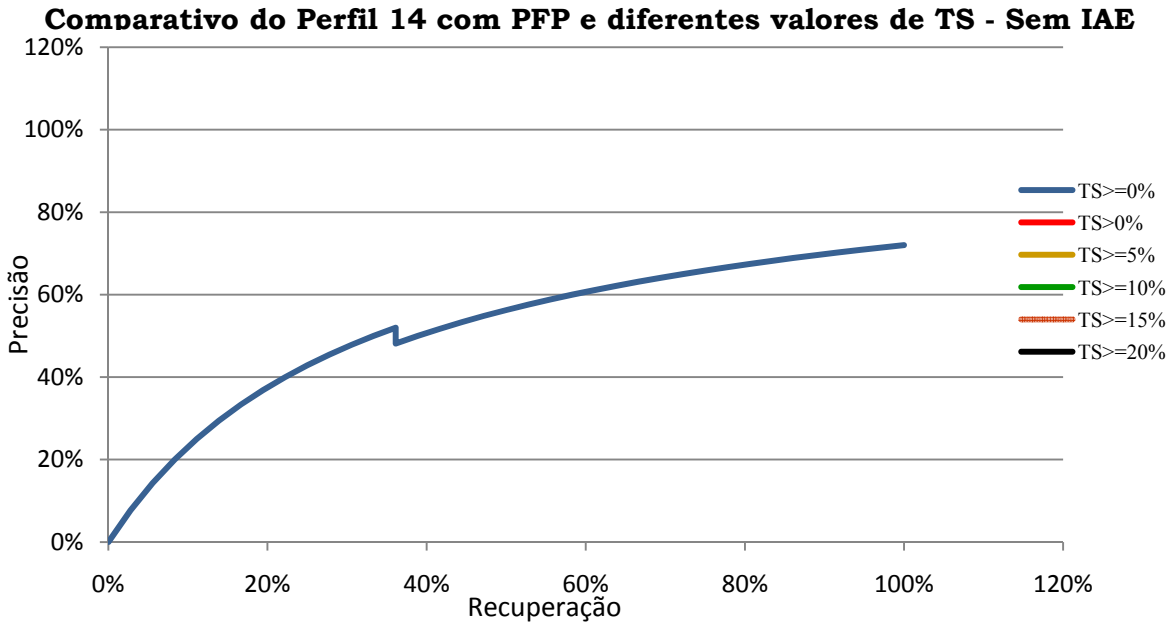


Figura 4.18 – Perfil 14: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

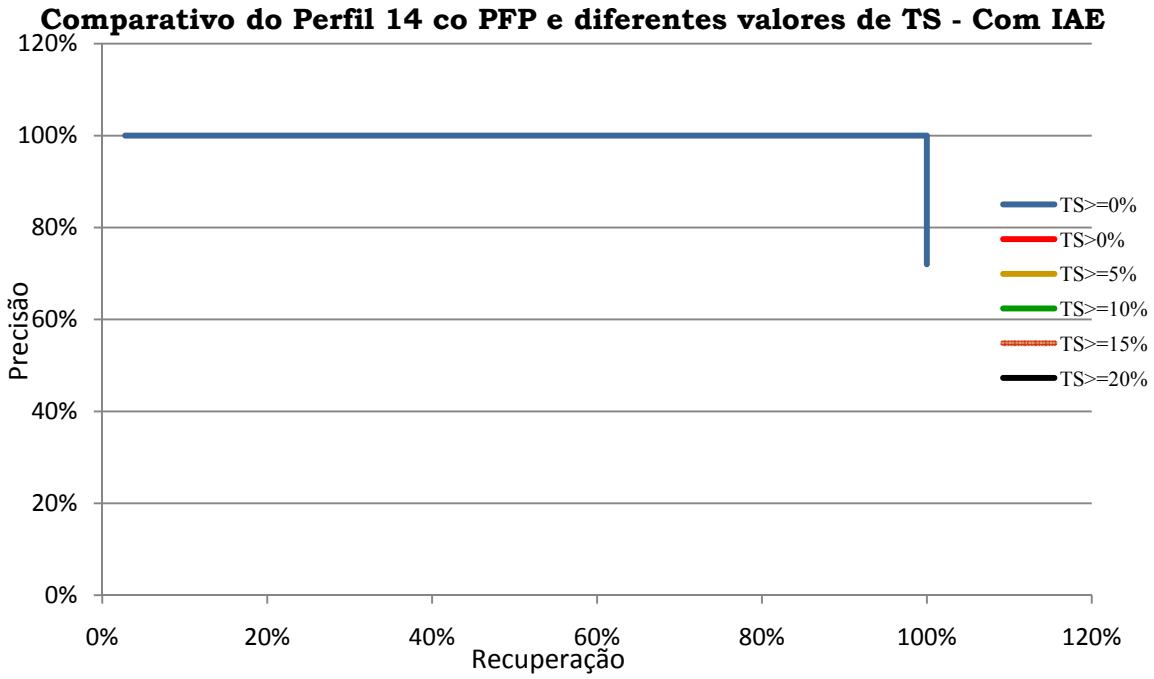


Figura 4.19 – Perfil 14: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

No teste anterior com o Perfil 4, não ocorreram resultados. Foi comprovado que o problema estava na escolha do número mínimo de palavras relevantes por publicação. Neste caso, foram encontrados resultados, mas pode-se observar que não similaridade entre os documentos. O grupo formado existe unicamente pelo fato de terem quantidades de palavras relevantes pequenas, o suficiente para ter valores de distância por *KNN* pequenos, sem representar similaridade com o resumo do perfil.

4.10.3 Perfil 15 – Assunto: *Business and Economics*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.41 representa a definição do perfil 15 com PFP e TS.

Tabela 4.41 – Perfil 15 com PFP e TS.

Assunto: ***Business and Economics***	
Título: ***Analyzing Mathematical Programs Using MProbe***	
Palavras Relevantes:	**Peso**
1ª) *und*	46%
2ª) *ifr*	21%
3ª) *angebot*	17%
4ª) *nachfrag*	17%
Resumo: Just as modern general-purpose programming languages (e.g., C++, Java) are supported by a suite of tools (debuggers, profilers, etc.), mathematical programming languages need supporting tools. MProbe is an example of a suite of tools supporting a mathematical programming language, in this case AMPL. MProbe includes tools for empirically estimating the shape of nonlinear functions of many variables, nonlinearly-constrained region shape, the effect of the objective shape on the ability to find a global optimum, tools for estimating the effectiveness of constraints and for navigating through the model, among others.	

As tabelas 4.42 e 4.43 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.20 e 4.21 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.42 – Comparativo das consultas do Perfil 15 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33%	50%	7%	50%	0%	0%	0%	0%	0%	0%	0%	0%
67%	67%	7%	33%	0%	0%	0%	0%	0%	0%	0%	0%
100%	75%	7%	25%	0%	0%	0%	0%	0%	0%	0%	0%
100%	60%	7%	20%	0%	0%	0%	0%	0%	0%	0%	0%
100%	50%	7%	17%	0%	0%	0%	0%	0%	0%	0%	0%
100%	43%	7%	14%	0%	0%	0%	0%	0%	0%	0%	0%
100%	38%	14%	25%	0%	0%	0%	0%	0%	0%	0%	0%
100%	33%	14%	22%	0%	0%	0%	0%	0%	0%	0%	0%
100%	30%	14%	20%	0%	0%	0%	0%	0%	0%	0%	0%
100%	27%	14%	18%	0%	0%	0%	0%	0%	0%	0%	0%
100%	25%	14%	17%	0%	0%	0%	0%	0%	0%	0%	0%
100%	23%	14%	15%	0%	0%	0%	0%	0%	0%	0%	0%
100%	21%	14%	14%	0%	0%	0%	0%	50%	7%	0%	0%
100%	20%	14%	13%	25%	7%	0%	0%	50%	7%	0%	0%
100%	19%	14%	13%	25%	6%	0%	0%	50%	6%	0%	0%
100%	18%	21%	18%	25%	6%	0%	0%	50%	6%	0%	0%
100%	17%	21%	17%	25%	6%	0%	0%	50%	6%	0%	0%
100%	16%	21%	16%	50%	11%	0%	0%	50%	5%	0%	0%
100%	15%	21%	15%	50%	10%	0%	0%	50%	5%	0%	0%
100%	14%	29%	19%	50%	10%	0%	0%	50%	5%	0%	0%
100%	14%	29%	18%	50%	9%	33%	5%	50%	5%	0%	0%
100%	13%	36%	22%	50%	9%	67%	9%	50%	4%	0%	0%
100%	13%	36%	21%	50%	8%	67%	8%	50%	4%	0%	0%
100%	12%	36%	20%	50%	8%	67%	8%	50%	4%	0%	0%
100%	12%	36%	19%	50%	8%	67%	8%	50%	4%	0%	0%
100%	11%	36%	19%	50%	7%	67%	7%	50%	4%	25%	4%
100%	11%	36%	18%	50%	7%	67%	7%	50%	4%	25%	4%
100%	10%	43%	21%	50%	7%	67%	7%	50%	3%	25%	3%
100%	10%	43%	20%	50%	7%	67%	7%	50%	3%	25%	3%
100%	10%	50%	23%	50%	6%	67%	6%	50%	3%	25%	3%
100%	9%	57%	25%	50%	6%	67%	6%	50%	3%	25%	3%
100%	9%	64%	27%	50%	6%	67%	6%	100%	6%	25%	3%
100%	9%	64%	26%	50%	6%	67%	6%	100%	6%	50%	6%
100%	9%	64%	26%	50%	6%	67%	6%	100%	6%	50%	6%
100%	8%	64%	25%	50%	6%	67%	6%	100%	6%	50%	6%
100%	8%	64%	24%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	64%	24%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	71%	26%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	71%	25%	75%	8%	67%	5%	100%	5%	50%	5%
100%	7%	71%	24%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	71%	24%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	79%	26%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	79%	25%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	86%	27%	75%	7%	67%	4%	100%	4%	75%	7%
100%	7%	93%	28%	75%	7%	67%	4%	100%	4%	75%	7%
100%	6%	100%	30%	100%	9%	67%	4%	100%	4%	75%	6%
100%	6%	100%	29%	100%	8%	67%	4%	100%	4%	100%	8%
100%	6%	100%	29%	100%	8%	100%	6%	100%	4%	100%	8%
100%	6%	100%	28%	100%	8%	100%	6%	100%	4%	100%	8%

Tabela 4.43 – Comparativo das consultas do Perfil 15 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
33%	50%	7%	50%	0%	0%	0%	0%	0%	0%	0%	0%
67%	67%	7%	33%	0%	0%	0%	0%	0%	0%	0%	0%
100%	75%	7%	25%	0%	0%	0%	0%	0%	0%	0%	0%
100%	60%	7%	20%	0%	0%	0%	0%	0%	0%	0%	0%
100%	50%	7%	17%	0%	0%	0%	0%	0%	0%	0%	0%
100%	43%	7%	14%	0%	0%	0%	0%	0%	0%	0%	0%
100%	38%	14%	25%	0%	0%	0%	0%	0%	0%	0%	0%
100%	33%	14%	22%	0%	0%	0%	0%	0%	0%	0%	0%
100%	30%	14%	20%	0%	0%	0%	0%	0%	0%	0%	0%
100%	27%	14%	18%	0%	0%	0%	0%	0%	0%	0%	0%
100%	25%	14%	17%	0%	0%	0%	0%	0%	0%	0%	0%
100%	23%	14%	15%	0%	0%	0%	0%	0%	0%	0%	0%
100%	21%	14%	14%	0%	0%	0%	0%	50%	7%	0%	0%
100%	20%	14%	13%	25%	7%	0%	0%	50%	7%	0%	0%
100%	19%	14%	13%	25%	6%	0%	0%	50%	6%	0%	0%
100%	18%	21%	18%	25%	6%	0%	0%	50%	6%	0%	0%
100%	17%	21%	17%	25%	6%	0%	0%	50%	6%	0%	0%
100%	16%	21%	16%	50%	11%	0%	0%	50%	5%	0%	0%
100%	15%	21%	15%	50%	10%	0%	0%	50%	5%	0%	0%
100%	14%	29%	19%	50%	10%	0%	0%	50%	5%	0%	0%
100%	14%	29%	18%	50%	9%	33%	5%	50%	5%	0%	0%
100%	13%	36%	22%	50%	9%	67%	9%	50%	4%	0%	0%
100%	13%	36%	21%	50%	8%	67%	8%	50%	4%	0%	0%
100%	12%	36%	20%	50%	8%	67%	8%	50%	4%	0%	0%
100%	12%	36%	19%	50%	8%	67%	8%	50%	4%	0%	0%
100%	11%	36%	19%	50%	7%	67%	7%	50%	4%	25%	4%
100%	11%	36%	18%	50%	7%	67%	7%	50%	4%	25%	4%
100%	10%	43%	21%	50%	7%	67%	7%	50%	3%	25%	3%
100%	10%	43%	20%	50%	7%	67%	7%	50%	3%	25%	3%
100%	10%	50%	23%	50%	6%	67%	6%	50%	3%	25%	3%
100%	9%	57%	25%	50%	6%	67%	6%	50%	3%	25%	3%
100%	9%	64%	27%	50%	6%	67%	6%	100%	6%	25%	3%
100%	9%	64%	26%	50%	6%	67%	6%	100%	6%	50%	6%
100%	9%	64%	26%	50%	6%	67%	6%	100%	6%	50%	6%
100%	8%	64%	25%	50%	6%	67%	6%	100%	6%	50%	6%
100%	8%	64%	24%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	64%	24%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	71%	26%	75%	8%	67%	5%	100%	5%	50%	5%
100%	8%	71%	25%	75%	8%	67%	5%	100%	5%	50%	5%
100%	7%	71%	24%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	71%	24%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	79%	26%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	79%	25%	75%	7%	67%	5%	100%	5%	50%	5%
100%	7%	86%	27%	75%	7%	67%	4%	100%	4%	75%	7%
100%	7%	93%	28%	75%	7%	67%	4%	100%	4%	75%	7%
100%	6%	100%	30%	100%	9%	67%	4%	100%	4%	75%	6%
100%	6%	100%	29%	100%	8%	67%	4%	100%	4%	100%	8%
100%	6%	100%	29%	100%	8%	100%	6%	100%	4%	100%	8%
100%	6%	100%	28%	100%	8%	100%	6%	100%	4%	100%	8%

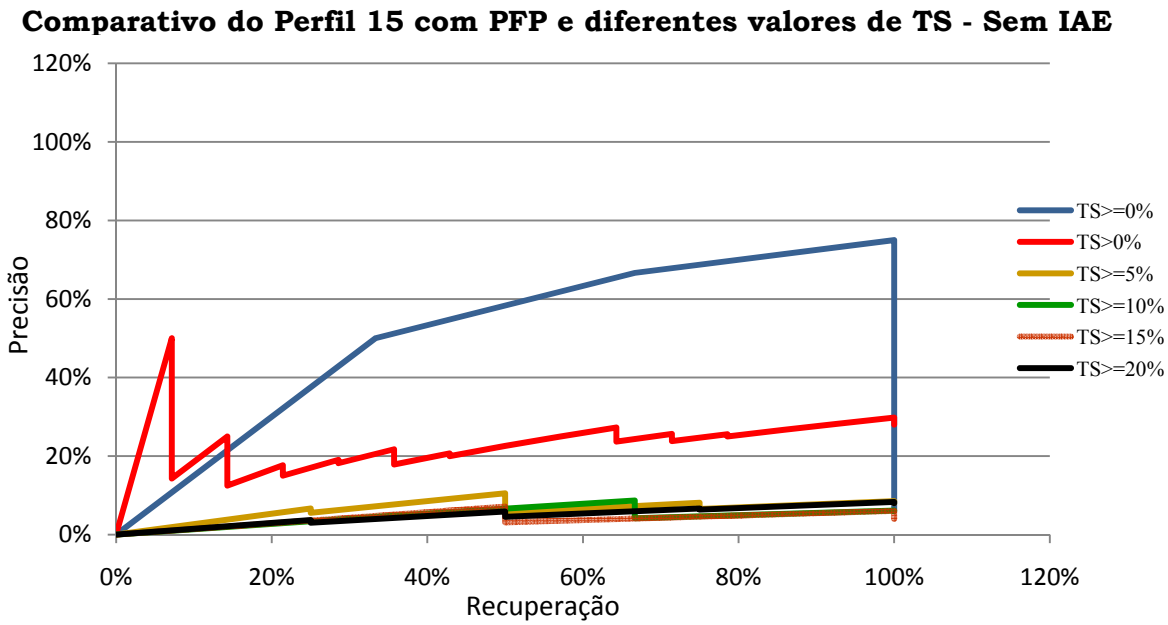


Figura 4.20 – Perfil 15: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

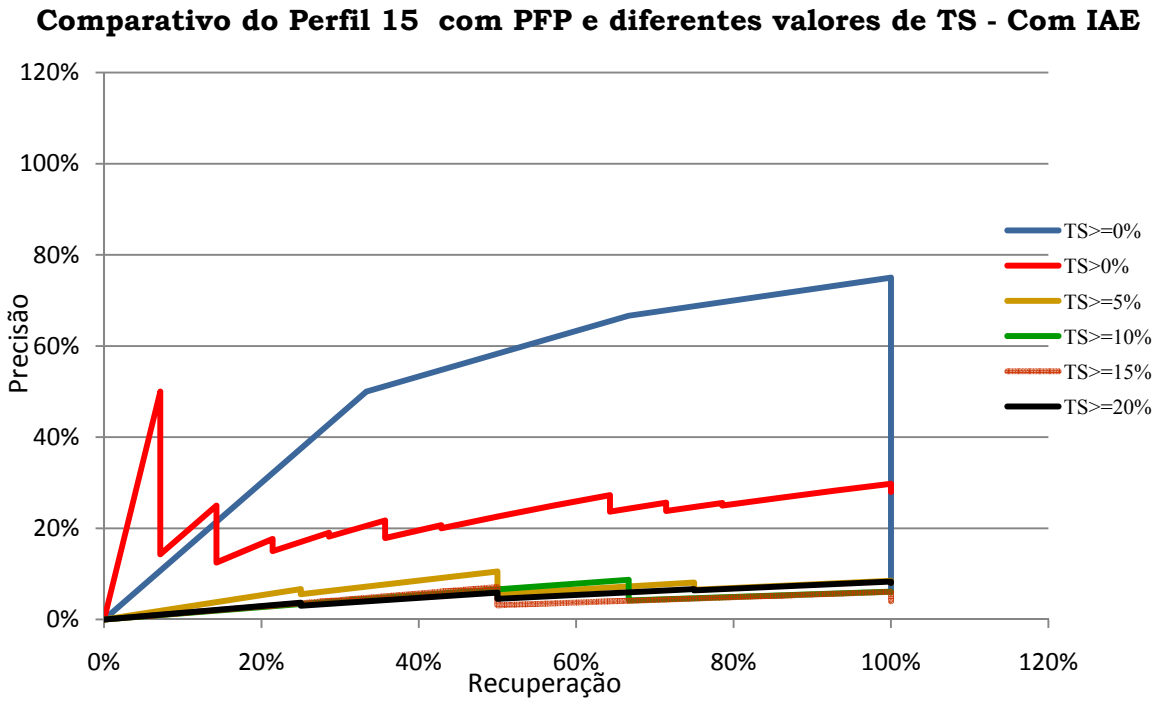


Figura 4.21 – Perfil 15: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Não houve diferenças entre o uso do IAE e sem o uso do mesmo. Indica que os termos escolhidos por PFP não foram adequados para este assunto.

4.10.4 Perfil 16 – Assunto: *Architeture and Design*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.44 representa a definição do perfil 16 com PFP e TS.

Tabela 4.44 – Perfil 16 com PFP e TS.

Assunto: ***Architetura and Design***	
Título: ***Maps and Geopolitics in Video Games***	
Palavras Relevantes:	**Peso**
1ª) *sich*	25%
2ª) *architect*	24%
3ª) *den*	23%
4ª) *build*	23%
Resumo: With the introduction of 3D-graphics space, a fundamental change in video game aesthetics and performativity took place: virtual space is now predominantly generated as presented space or spatial presence (Poole 2000/McMahan 2003). However, spatial formations in video games confusingly are still referred to as "representational space" (Wolf 1997) in general, even though the pictorial appearance of the game is not always a literal representation. Rather, representations in video games have specific functions that define their role in the interactive play just like any particular form of spatiality does (Aarseth 2001/Fernández-Vara et al. 2005/Taylor 2005). The function of representations in games can sufficiently be described by Henri Lefebvre's (1991) trialectic of spatial processes, according to which the individual's perception of spacetime and the spatiotemporal structures of the social — "representational spaces" in Lefebvre's words — are reciprocally mediated by representations of space — namely, maps. Space in Lefebvre's understanding is thus threefold: a combination of perceived (perçu), conceived (conçu) and lived (vécu). Video games today are mostly a presentation of perceptual space in the way Lefebvre addresses the individual experience of space or what he also calls the "spatial practice." In contrast, representations of space differ from this phenomenal experience of space: as they are in real-life contexts, video game maps are essential for orientation, especially in games played from the first-person perspective, for in those games, one not only needs to see what one is aiming at, but also where one is located within the entire setting of the game. For this reason, maps in video games are either fully displayed and function as representations of the whole "playground" (which gamers usually call the "map"), or they are reduced to a visual element within the display, most frequently a radar that allows for orientation within the periphery of the position of the avatar or the ego in play.	

As tabelas 4.45 e 4.46 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.22 e 4.23 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.45 – Comparativo das consultas do Perfil 16 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
33%	100%	33%	100%	4%	100%	4%	100%	13%	100%	100%	100%
33%	50%	33%	50%	8%	100%	9%	100%	25%	100%	100%	50%
33%	33%	33%	33%	12%	100%	13%	100%	38%	100%	100%	33%
33%	25%	33%	25%	16%	100%	17%	100%	50%	100%	100%	25%
33%	20%	33%	20%	20%	100%	22%	100%	63%	100%	100%	20%
33%	17%	33%	17%	20%	83%	26%	100%	75%	100%	100%	17%
33%	14%	33%	14%	24%	86%	30%	100%	88%	100%	100%	14%
33%	13%	33%	13%	24%	75%	35%	100%	88%	88%	100%	13%
33%	11%	33%	11%	24%	67%	39%	100%	88%	78%	100%	11%
33%	10%	33%	10%	24%	60%	43%	100%	88%	70%	100%	10%
33%	9%	33%	9%	24%	55%	43%	91%	88%	64%	100%	9%
33%	8%	33%	8%	28%	58%	43%	83%	88%	58%	100%	8%
33%	8%	33%	8%	32%	62%	48%	85%	88%	54%	100%	8%
33%	7%	33%	7%	32%	57%	52%	86%	88%	50%	100%	7%
33%	7%	33%	7%	32%	53%	57%	87%	88%	47%	100%	7%
33%	6%	33%	6%	36%	56%	57%	81%	88%	44%	100%	6%
33%	6%	33%	6%	40%	59%	61%	82%	88%	41%		
33%	6%	33%	6%	40%	56%	61%	78%	88%	39%		
33%	5%	33%	5%	44%	58%	65%	79%	88%	37%		
33%	5%	33%	5%	48%	60%	65%	75%	100%	40%		
33%	5%	33%	5%	52%	62%	65%	71%	100%	38%		
33%	5%	33%	5%	52%	59%	65%	68%	100%	36%		
33%	4%	33%	4%	56%	61%	70%	70%	100%	35%		
33%	4%	67%	8%	56%	58%	70%	67%	100%	33%		
33%	4%	67%	8%	56%	56%	74%	68%	100%	32%		
33%	4%	67%	8%	60%	58%	74%	65%	100%	31%		
33%	4%	67%	7%	64%	59%	74%	63%	100%	30%		
33%	4%	100%	11%	64%	57%	74%	61%	100%	29%		
33%	3%	100%	10%	64%	55%	74%	59%	100%	28%		
33%	3%	100%	10%	68%	57%	74%	57%	100%	27%		
33%	3%	100%	10%	68%	55%	74%	55%	100%	26%		
33%	3%	100%	9%	68%	53%	74%	53%	100%	25%		
33%	3%	100%	9%	68%	52%	78%	55%	100%	24%		
33%	3%	100%	9%	68%	50%	78%	53%	100%	24%		
33%	3%	100%	9%	68%	49%	83%	54%	100%	23%		
33%	3%	100%	8%	72%	50%	87%	56%	100%	22%		
67%	5%	100%	8%	76%	51%	87%	54%	100%	22%		
67%	5%	100%	8%	76%	50%	87%	53%	100%	21%		
67%	5%	100%	8%	76%	49%	87%	51%	100%	21%		
67%	5%	100%	8%	80%	50%	87%	50%	100%	20%		
100%	7%	100%	7%	84%	51%	91%	51%	100%	20%		
100%	7%	100%	7%	84%	50%	91%	50%	100%	19%		
100%	7%	100%	7%	84%	49%	91%	49%	100%	19%		
100%	7%	100%	7%	84%	48%	91%	48%	100%	18%		
100%	7%	100%	7%	88%	49%	91%	47%	100%	18%		
100%	7%	100%	7%	92%	50%	91%	46%	100%	17%		
100%	6%	100%	6%	92%	49%	91%	45%	100%	17%		
100%	6%	100%	6%	92%	48%	96%	46%	100%	17%		
100%	6%	100%	6%	96%	49%	100%	47%	100%	16%		
100%	6%	100%	6%	100%	50%	100%	46%	100%	16%		

Tabela 4.46 – Comparativo das consultas do Perfil 16 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
33%	100%	33%	100%	4%	100%	4%	100%	13%	100%	0%	0%
33%	50%	33%	50%	8%	100%	9%	100%	13%	50%	0%	0%
33%	33%	33%	33%	8%	67%	13%	100%	13%	33%	0%	0%
33%	25%	33%	25%	12%	75%	17%	100%	13%	25%	0%	0%
33%	20%	33%	20%	16%	80%	22%	100%	13%	20%	0%	0%
33%	17%	33%	17%	20%	83%	26%	100%	13%	17%	100%	17%
33%	14%	33%	14%	24%	86%	30%	100%	25%	29%	100%	14%
33%	13%	33%	13%	28%	88%	35%	100%	38%	38%	100%	13%
33%	11%	33%	11%	28%	78%	39%	100%	50%	44%	100%	11%
33%	10%	33%	10%	28%	70%	43%	100%	63%	50%	100%	10%
33%	9%	33%	9%	28%	64%	48%	100%	75%	55%	100%	9%
33%	8%	33%	8%	28%	58%	52%	100%	88%	58%	100%	8%
33%	8%	33%	8%	32%	62%	52%	92%	100%	62%	100%	8%
33%	7%	33%	7%	36%	64%	52%	86%	100%	57%	100%	7%
33%	7%	33%	7%	36%	60%	57%	87%	100%	53%	100%	7%
33%	6%	33%	6%	36%	56%	61%	88%	100%	50%	100%	6%
33%	6%	33%	6%	40%	59%	65%	88%	100%	47%		
33%	6%	33%	6%	44%	61%	65%	83%	100%	44%		
33%	5%	33%	5%	44%	58%	70%	84%	100%	42%		
33%	5%	33%	5%	48%	60%	70%	80%	100%	40%		
33%	5%	33%	5%	52%	62%	74%	81%	100%	38%		
33%	5%	33%	5%	56%	64%	74%	77%	100%	36%		
33%	4%	33%	4%	56%	61%	74%	74%	100%	35%		
33%	4%	67%	8%	60%	63%	74%	71%	100%	33%		
33%	4%	67%	8%	60%	60%	78%	72%	100%	32%		
33%	4%	67%	8%	60%	58%	78%	69%	100%	31%		
33%	4%	67%	7%	64%	59%	78%	67%	100%	30%		
33%	4%	100%	11%	64%	57%	78%	64%	100%	29%		
33%	3%	100%	10%	64%	55%	78%	62%	100%	28%		
33%	3%	100%	10%	68%	57%	78%	60%	100%	27%		
33%	3%	100%	10%	68%	55%	78%	58%	100%	26%		
33%	3%	100%	9%	68%	53%	78%	56%	100%	25%		
33%	3%	100%	9%	68%	52%	78%	55%	100%	24%		
33%	3%	100%	9%	68%	50%	83%	56%	100%	24%		
33%	3%	100%	9%	68%	49%	83%	54%	100%	23%		
33%	3%	100%	8%	72%	50%	87%	56%	100%	22%		
67%	5%	100%	8%	76%	51%	91%	57%	100%	22%		
67%	5%	100%	8%	76%	50%	91%	55%	100%	21%		
67%	5%	100%	8%	76%	49%	91%	54%	100%	21%		
67%	5%	100%	8%	80%	50%	91%	53%	100%	20%		
100%	7%	100%	7%	84%	51%	91%	51%	100%	20%		
100%	7%	100%	7%	84%	50%	96%	52%	100%	19%		
100%	7%	100%	7%	84%	49%	96%	51%	100%	19%		
100%	7%	100%	7%	84%	48%	96%	50%	100%	18%		
100%	7%	100%	7%	88%	49%	96%	49%	100%	18%		
100%	7%	100%	7%	92%	50%	96%	48%	100%	17%		
100%	6%	100%	6%	92%	49%	96%	47%	100%	17%		
100%	6%	100%	6%	92%	48%	96%	46%	100%	17%		
100%	6%	100%	6%	96%	49%	100%	47%	100%	16%		
100%	6%	100%	6%	100%	50%	100%	46%	100%	16%		

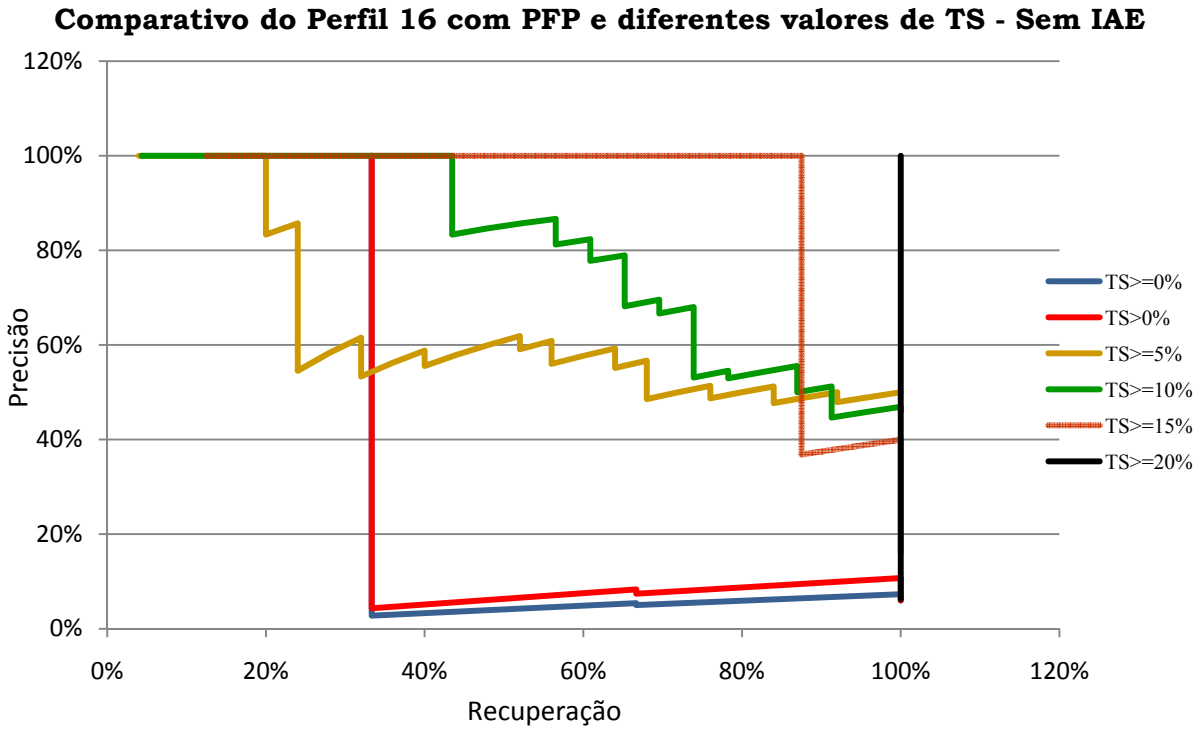


Figura 4.22 – Perfil 16: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

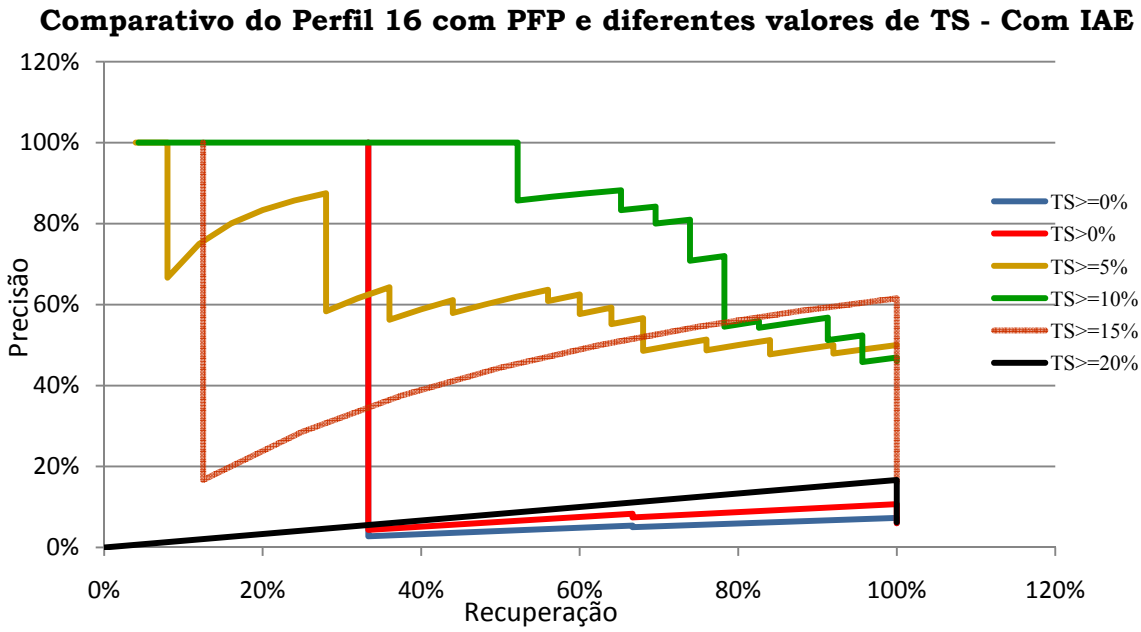


Figura 4.23 – Perfil 16: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Note que sem o uso do IAE, os melhores resultados foram obtidos sobre as consultas com 15% ou mais de Taxa de Similaridade, mas com o emprego do IAE os melhores resultados foram obtidos utilizando uma Taxa de Similaridade bem menor, de 10% ou mais e de 5% ou mais. A única perda significativa de precisão ocorreu quando usada a Taxa de Similaridade de 20% ou mais.

4.10.5 Perfil 17 – Assunto: *Professional and Applied Computing*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.47 representa a definição do perfil 17 com PFP e TS.

Tabela 4.47 – Perfil 17 com PFP e TS.

Assunto: ***Professional and Applied Computing***	
Título: ***Zone Files and Resource Records***	
Palavras Relevantes:	**Peso**
1ª) *ll*	29%
2ª) *applic*	26%
3ª) *web*	24%
4ª) *server*	21%
Resumo: This chapter described the format and content of zone files. You learned about the $TTL directive, used to set the default TTL for the zone. You also encountered the $ORIGIN directive, used to set the base name for the zone, and the $ORIGIN substitution rule, the cause of much DNS aggravation. Using the example zone file as a guide, the text explained the various Resource Record types used to construct basic zone files such as the Start of Authority, Name Server, Mail Exchanger, and Address Resource Records. Chapter 3 explains DNS operations: the types of DNS queries that may be used; reverse mapping, the process by which an IP address may be mapped to a host name; zone transfers, the method by which zone files are updated from the master to the slave name servers; and finally, a brief overview of the security issues involved in running a DNS service.	

As tabelas 4.48 e 4.49 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.24 e 4.25 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.48 – Comparativo das consultas do Perfil 17 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
25%	100%	6%	100%	3%	100%	2%	100%	2%	100%	3%	100%
25%	50%	6%	50%	5%	100%	4%	100%	5%	100%	6%	100%
25%	33%	12%	67%	8%	100%	6%	100%	7%	100%	9%	100%
50%	50%	18%	75%	10%	100%	9%	100%	10%	100%	9%	75%
75%	60%	18%	60%	13%	100%	11%	100%	12%	100%	12%	80%
75%	50%	18%	50%	15%	100%	13%	100%	15%	100%	15%	83%
75%	43%	18%	43%	15%	86%	15%	100%	17%	100%	18%	86%
75%	38%	18%	38%	15%	75%	17%	100%	20%	100%	21%	88%
75%	33%	18%	33%	15%	67%	19%	100%	22%	100%	24%	89%
75%	30%	18%	30%	18%	70%	21%	100%	24%	100%	27%	90%
75%	27%	24%	36%	20%	73%	23%	100%	27%	100%	30%	91%
75%	25%	24%	33%	23%	75%	26%	100%	27%	92%	33%	92%
75%	23%	24%	31%	25%	77%	28%	100%	29%	92%	36%	92%
75%	21%	29%	36%	28%	79%	30%	100%	29%	86%	39%	93%
75%	20%	29%	33%	30%	80%	32%	100%	32%	87%	42%	93%
75%	19%	29%	31%	33%	81%	34%	100%	34%	88%	45%	94%
75%	18%	35%	35%	35%	82%	36%	100%	37%	88%	48%	94%
75%	17%	41%	39%	38%	83%	38%	100%	39%	89%	52%	94%
100%	21%	47%	42%	40%	84%	40%	100%	41%	89%	52%	89%
100%	20%	47%	40%	43%	85%	43%	100%	41%	85%	55%	90%
100%	19%	47%	38%	45%	86%	43%	95%	41%	81%	58%	90%
100%	18%	47%	36%	48%	86%	45%	95%	44%	82%	58%	86%
100%	17%	47%	35%	50%	87%	47%	96%	46%	83%	58%	83%
100%	17%	47%	33%	53%	88%	49%	96%	49%	83%	58%	79%
100%	16%	47%	32%	55%	88%	51%	96%	51%	84%	61%	80%
100%	15%	47%	31%	55%	85%	53%	96%	54%	85%	64%	81%
100%	15%	47%	30%	55%	81%	55%	96%	56%	85%	67%	81%
100%	14%	47%	29%	55%	79%	57%	96%	59%	86%	70%	82%
100%	14%	47%	28%	58%	79%	60%	97%	61%	86%	70%	79%
100%	13%	47%	27%	60%	80%	62%	97%	61%	83%	70%	77%
100%	13%	53%	29%	63%	81%	64%	97%	63%	84%	73%	77%
100%	13%	59%	31%	65%	81%	66%	97%	66%	84%	76%	78%
100%	12%	65%	33%	68%	82%	66%	94%	66%	82%	79%	79%
100%	12%	71%	35%	70%	82%	68%	94%	68%	82%	82%	79%
100%	11%	71%	34%	73%	83%	70%	94%	71%	83%	82%	77%
100%	11%	71%	33%	75%	83%	72%	94%	73%	83%	82%	75%
100%	11%	71%	32%	75%	81%	74%	95%	76%	84%	85%	76%
100%	11%	71%	32%	78%	82%	77%	95%	78%	84%	85%	74%
100%	10%	71%	31%	80%	82%	79%	95%	78%	82%	85%	72%
100%	10%	71%	30%	83%	83%	81%	95%	80%	83%	88%	73%
100%	10%	71%	29%	83%	80%	83%	95%	83%	83%	91%	73%
100%	10%	71%	29%	83%	79%	85%	95%	85%	83%	94%	74%
100%	9%	76%	30%	85%	79%	87%	95%	88%	84%	94%	72%
100%	9%	82%	32%	88%	80%	89%	95%	88%	82%	94%	70%
100%	9%	88%	33%	88%	78%	91%	96%	90%	82%	94%	69%
100%	9%	94%	35%	90%	78%	94%	96%	90%	80%	94%	67%
100%	9%	94%	34%	93%	79%	96%	96%	93%	81%	97%	68%
100%	8%	100%	35%	95%	79%	98%	96%	95%	81%	100%	69%
100%	8%	100%	35%	98%	80%	98%	94%	98%	82%	100%	67%
100%	8%	100%	34%	100%	80%	100%	94%	100%	82%	100%	66%

Tabela 4.49 – Comparativo das consultas do Perfil 17 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
33%	100%	33%	100%	4%	100%	4%	100%	13%	100%	0%	0%
33%	50%	33%	50%	8%	100%	9%	100%	13%	50%	0%	0%
33%	33%	33%	33%	8%	67%	13%	100%	13%	33%	0%	0%
33%	25%	33%	25%	12%	75%	17%	100%	13%	25%	0%	0%
33%	20%	33%	20%	16%	80%	22%	100%	13%	20%	0%	0%
33%	17%	33%	17%	20%	83%	26%	100%	13%	17%	100%	17%
33%	14%	33%	14%	24%	86%	30%	100%	25%	29%	100%	14%
33%	13%	33%	13%	28%	88%	35%	100%	38%	38%	100%	13%
33%	11%	33%	11%	28%	78%	39%	100%	50%	44%	100%	11%
33%	10%	33%	10%	28%	70%	43%	100%	63%	50%	100%	10%
33%	9%	33%	9%	28%	64%	48%	100%	75%	55%	100%	9%
33%	8%	33%	8%	28%	58%	52%	100%	88%	58%	100%	8%
33%	8%	33%	8%	32%	62%	52%	92%	100%	62%	100%	8%
33%	7%	33%	7%	36%	64%	52%	86%	100%	57%	100%	7%
33%	7%	33%	7%	36%	60%	57%	87%	100%	53%	100%	7%
33%	6%	33%	6%	36%	56%	61%	88%	100%	50%	100%	6%
33%	6%	33%	6%	40%	59%	65%	88%	100%	47%		
33%	6%	33%	6%	44%	61%	65%	83%	100%	44%		
33%	5%	33%	5%	44%	58%	70%	84%	100%	42%		
33%	5%	33%	5%	48%	60%	70%	80%	100%	40%		
33%	5%	33%	5%	52%	62%	74%	81%	100%	38%		
33%	5%	33%	5%	56%	64%	74%	77%	100%	36%		
33%	4%	33%	4%	56%	61%	74%	74%	100%	35%		
33%	4%	67%	8%	60%	63%	74%	71%	100%	33%		
33%	4%	67%	8%	60%	60%	78%	72%	100%	32%		
33%	4%	67%	8%	60%	58%	78%	69%	100%	31%		
33%	4%	67%	7%	64%	59%	78%	67%	100%	30%		
33%	4%	100%	11%	64%	57%	78%	64%	100%	29%		
33%	3%	100%	10%	64%	55%	78%	62%	100%	28%		
33%	3%	100%	10%	68%	57%	78%	60%	100%	27%		
33%	3%	100%	10%	68%	55%	78%	58%	100%	26%		
33%	3%	100%	9%	68%	53%	78%	56%	100%	25%		
33%	3%	100%	9%	68%	52%	78%	55%	100%	24%		
33%	3%	100%	9%	68%	50%	83%	56%	100%	24%		
33%	3%	100%	9%	68%	49%	83%	54%	100%	23%		
33%	3%	100%	8%	72%	50%	87%	56%	100%	22%		
67%	5%	100%	8%	76%	51%	91%	57%	100%	22%		
67%	5%	100%	8%	76%	50%	91%	55%	100%	21%		
67%	5%	100%	8%	76%	49%	91%	54%	100%	21%		
67%	5%	100%	8%	80%	50%	91%	53%	100%	20%		
100%	7%	100%	7%	84%	51%	91%	51%	100%	20%		
100%	7%	100%	7%	84%	50%	96%	52%	100%	19%		
100%	7%	100%	7%	84%	49%	96%	51%	100%	19%		
100%	7%	100%	7%	84%	48%	96%	50%	100%	18%		
100%	7%	100%	7%	88%	49%	96%	49%	100%	18%		
100%	7%	100%	7%	92%	50%	96%	48%	100%	17%		
100%	6%	100%	6%	92%	49%	96%	47%	100%	17%		
100%	6%	100%	6%	92%	48%	96%	46%	100%	17%		
100%	6%	100%	6%	96%	49%	100%	47%	100%	16%		
100%	6%	100%	6%	100%	50%	100%	46%	100%	16%		

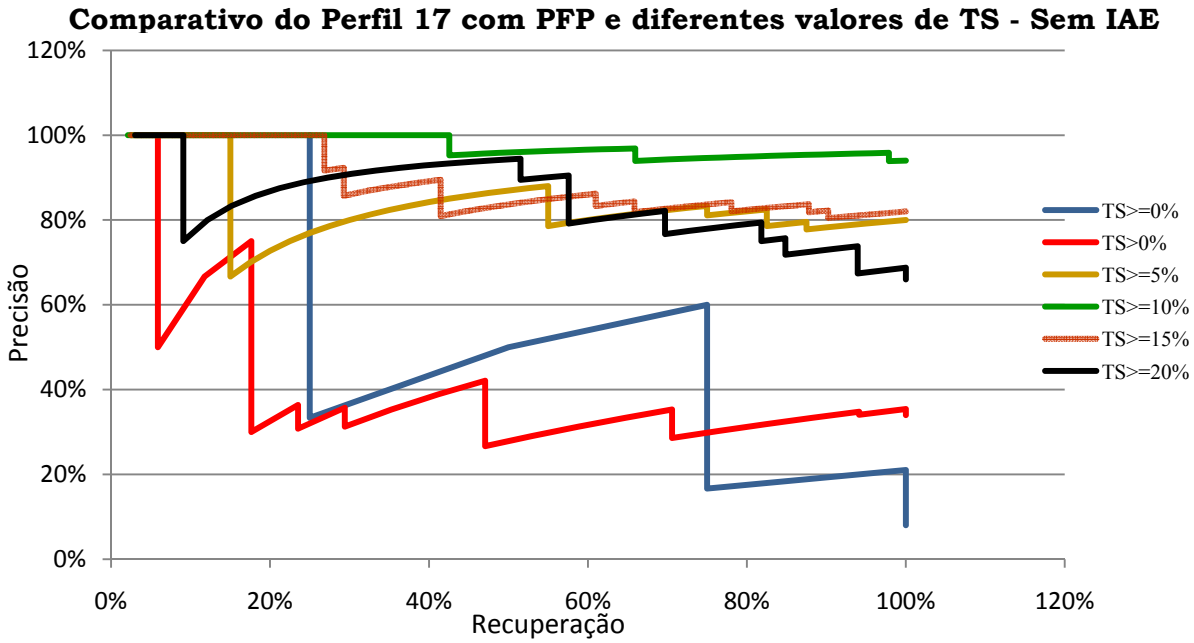


Figura 4.24 – Perfil 17: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

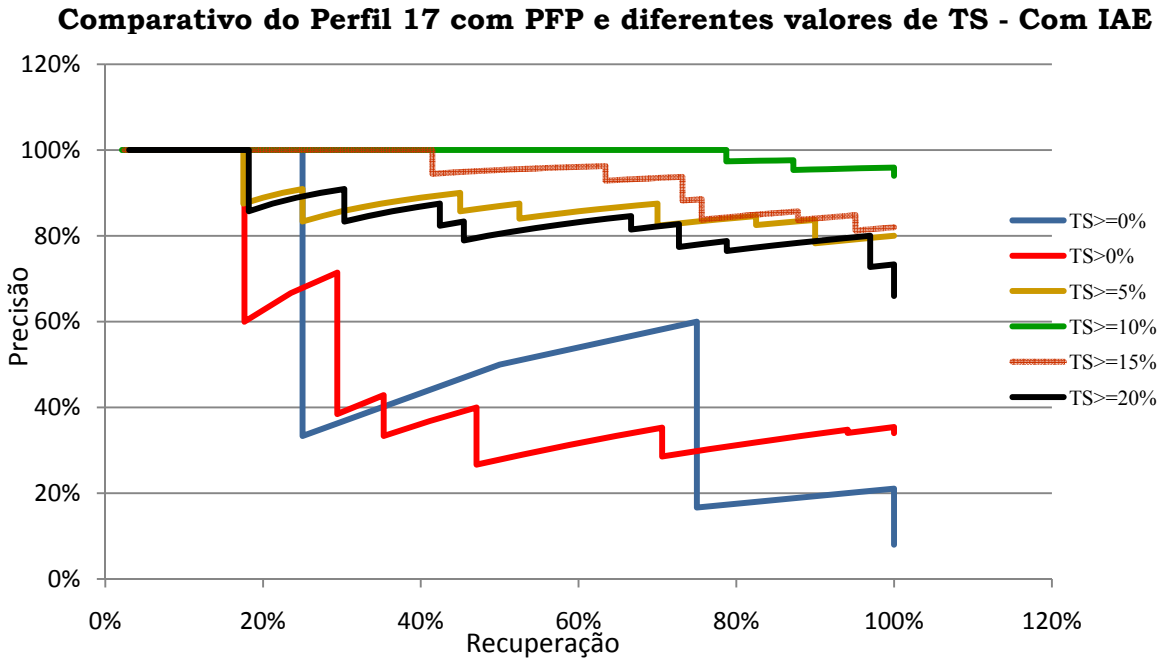


Figura 4.25 – Perfil 17: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Neste caso, observa-se que o IAE obteve melhorias em todos os casos, sendo que apenas a Taxa de Similaridade de 0% ou maior se manteve estável.

4.10.6 Perfil 18 – Assunto: *Medicine*

Utilizando de Palavras Freqüentes e Preditivas e Taxa de Similaridade. A tabela 4.50 representa a definição do perfil 18 com PFP e TS.

Tabela 4.50 – Perfil 18 com PFP e TS.

Assunto: ***Medicine***	
Título: ***Epigastrica inferior Atypische Blutungsquelle bei Beckenfraktur***	
Palavras Relevantes:	**Peso**
1ª) *menschen*	30%
2ª) *der*	29%
3ª) *la*	21%
4ª) *sympathectomi*	20%
Resumo: The most common mechanism of pelvic fractures after blunt trauma is lateral commpression of the pelvis. Most of these fractures are of slight severity but it is possible, that a life-threatening hemorrhage can appear. The inferior epigastric artery is an atypical bleedingsite but it has to be considered in the search of the origin of the hemorrhage. Diagnostic tools are ultasound, computed tomography and angiography. After external fixation of the pelvis and persistent haemodynamic active bleeding is the percutaneous transcatheter embolization (PTE) in our opinion the treatment of choice. The introduced case is confirm with this statement.	

As tabelas 4.51 e 4.52 apresentam os resultados das consultas Sem IAE e Com IAE respectivamente e as figuras 4.26 e 4.27 apresentam os gráficos comparativos entre as diferentes consultas Sem IAE e Com IAE respectivamente.

Tabela 4.51 – Comparativo das consultas do Perfil 18 com PFP e TS Sem IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
SEM IAE		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**		**SEM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
25%	100%	6%	100%	3%	100%	2%	100%	2%	100%	3%	100%
25%	50%	12%	100%	5%	100%	4%	100%	5%	100%	6%	100%
25%	33%	18%	100%	8%	100%	6%	100%	7%	100%	9%	100%
50%	50%	18%	75%	10%	100%	9%	100%	10%	100%	12%	100%
75%	60%	18%	60%	13%	100%	11%	100%	12%	100%	15%	100%
75%	50%	24%	67%	15%	100%	13%	100%	15%	100%	18%	100%
75%	43%	29%	71%	18%	100%	15%	100%	17%	100%	18%	86%
75%	38%	29%	63%	18%	88%	17%	100%	20%	100%	21%	88%
75%	33%	29%	56%	20%	89%	19%	100%	22%	100%	24%	89%
75%	30%	29%	50%	23%	90%	21%	100%	24%	100%	27%	90%
75%	27%	29%	45%	25%	91%	23%	100%	27%	100%	30%	91%
75%	25%	29%	42%	25%	83%	26%	100%	29%	100%	30%	83%
75%	23%	29%	38%	28%	85%	28%	100%	32%	100%	33%	85%
75%	21%	35%	43%	30%	86%	30%	100%	34%	100%	36%	86%
75%	20%	35%	40%	33%	87%	32%	100%	37%	100%	39%	87%
75%	19%	35%	38%	35%	88%	34%	100%	39%	100%	42%	88%
75%	18%	35%	35%	38%	88%	36%	100%	41%	100%	42%	82%
75%	17%	35%	33%	40%	89%	38%	100%	41%	94%	45%	83%
100%	21%	41%	37%	43%	89%	40%	100%	44%	95%	45%	79%
100%	20%	47%	40%	45%	90%	43%	100%	46%	95%	48%	80%
100%	19%	47%	38%	45%	86%	45%	100%	49%	95%	52%	81%
100%	18%	47%	36%	48%	86%	47%	100%	51%	95%	55%	82%
100%	17%	47%	35%	50%	87%	49%	100%	54%	96%	58%	83%
100%	17%	47%	33%	53%	88%	51%	100%	56%	96%	61%	83%
100%	16%	47%	32%	53%	84%	53%	100%	59%	96%	64%	84%
100%	15%	47%	31%	55%	85%	55%	100%	61%	96%	67%	85%
100%	15%	47%	30%	58%	85%	57%	100%	63%	96%	67%	81%
100%	14%	47%	29%	60%	86%	60%	100%	63%	93%	70%	82%
100%	14%	47%	28%	63%	86%	62%	100%	66%	93%	73%	83%
100%	13%	47%	27%	65%	87%	64%	100%	68%	93%	73%	80%
100%	13%	53%	29%	68%	87%	66%	100%	71%	94%	73%	77%
100%	13%	59%	31%	70%	88%	68%	100%	73%	94%	76%	78%
100%	12%	65%	33%	70%	85%	70%	100%	73%	91%	79%	79%
100%	12%	71%	35%	70%	82%	72%	100%	73%	88%	79%	76%
100%	11%	71%	34%	73%	83%	74%	100%	76%	89%	82%	77%
100%	11%	71%	33%	75%	83%	77%	100%	76%	86%	85%	78%
100%	11%	71%	32%	78%	84%	79%	100%	76%	84%	88%	78%
100%	11%	71%	32%	80%	84%	79%	97%	78%	84%	91%	79%
100%	10%	71%	31%	83%	85%	81%	97%	80%	85%	94%	79%
100%	10%	71%	30%	83%	83%	83%	98%	83%	85%	97%	80%
100%	10%	71%	29%	85%	83%	85%	98%	85%	85%	97%	78%
100%	10%	71%	29%	88%	83%	87%	98%	88%	86%	97%	76%
100%	9%	76%	30%	90%	84%	87%	95%	88%	84%	97%	74%
100%	9%	82%	32%	90%	82%	89%	95%	90%	84%	97%	73%
100%	9%	88%	33%	90%	80%	91%	96%	93%	84%	100%	73%
100%	9%	94%	35%	90%	78%	94%	96%	95%	85%	100%	72%
100%	9%	94%	34%	93%	79%	96%	96%	95%	83%	100%	70%
100%	8%	100%	35%	95%	79%	98%	96%	95%	81%	100%	69%
100%	8%	100%	35%	98%	80%	100%	96%	98%	82%	100%	67%
100%	8%	100%	34%	100%	80%	100%	94%	100%	82%	100%	66%

Tabela 4.52 – Comparativo das consultas do Perfil 18 com PFP e TS Com IAE.

TS >= 0%		*TS > 0%*		*TS >=5%*		*TS >=10%*		*TS >= 15%*		*TS>=20%*	
COM IAE		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**		**COM IAE**	
Recup.	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**	**Recup.**	**Precisão**
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	33%	50%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	33%	33%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	67%	50%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	67%	40%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	67%	33%	0%	0%	0%	0%	0%	0%
0%	0%	0%	0%	67%	29%	0%	0%	33%	14%	0%	0%
0%	0%	0%	0%	67%	25%	0%	0%	33%	13%	20%	13%
0%	0%	0%	0%	67%	22%	0%	0%	33%	11%	20%	11%
0%	0%	0%	0%	67%	20%	0%	0%	33%	10%	20%	10%
0%	0%	0%	0%	67%	18%	0%	0%	33%	9%	20%	9%
0%	0%	0%	0%	67%	17%	0%	0%	33%	8%	40%	17%
0%	0%	0%	0%	67%	15%	50%	8%	67%	15%	40%	15%
0%	0%	0%	0%	67%	14%	100%	14%	67%	14%	40%	14%
0%	0%	0%	0%	67%	13%	100%	13%	67%	13%	40%	13%
0%	0%	0%	0%	67%	13%	100%	13%	67%	13%	60%	19%
0%	0%	0%	0%	67%	12%	100%	12%	67%	12%	60%	18%
0%	0%	0%	0%	67%	11%	100%	11%	67%	11%	60%	17%
0%	0%	0%	0%	67%	11%	100%	11%	67%	11%	60%	16%
0%	0%	0%	0%	67%	10%	100%	10%	67%	10%	80%	20%
0%	0%	0%	0%	67%	10%	100%	10%	67%	10%	80%	19%
0%	0%	0%	0%	67%	9%	100%	9%	67%	9%	80%	18%
0%	0%	100%	4%	67%	9%	100%	9%	67%	9%	80%	17%
0%	0%	100%	4%	67%	8%	100%	8%	67%	8%	100%	21%
0%	0%	100%	4%	67%	8%	100%	8%	67%	8%	100%	20%
0%	0%	100%	4%	67%	8%	100%	8%	67%	8%	100%	19%
0%	0%	100%	4%	67%	7%	100%	7%	67%	7%	100%	19%
0%	0%	100%	4%	67%	7%	100%	7%	67%	7%	100%	18%
0%	0%	100%	3%	67%	7%	100%	7%	67%	7%	100%	17%
0%	0%	100%	3%	67%	7%	100%	7%	67%	7%	100%	17%
0%	0%	100%	3%	67%	6%	100%	6%	67%	6%	100%	16%
0%	0%	100%	3%	100%	9%	100%	6%	67%	6%	100%	16%
0%	0%	100%	3%	100%	9%	100%	6%	67%	6%	100%	15%
0%	0%	100%	3%	100%	9%	100%	6%	67%	6%	100%	15%
0%	0%	100%	3%	100%	9%	100%	6%	67%	6%	100%	14%
0%	0%	100%	3%	100%	8%	100%	6%	67%	6%	100%	14%
0%	0%	100%	3%	100%	8%	100%	5%	67%	5%	100%	14%
0%	0%	100%	3%	100%	8%	100%	5%	67%	5%	100%	13%
0%	0%	100%	3%	100%	8%	100%	5%	67%	5%	100%	13%
0%	0%	100%	3%	100%	8%	100%	5%	67%	5%	100%	13%
0%	0%	100%	2%	100%	7%	100%	5%	67%	5%	100%	12%
0%	0%	100%	2%	100%	7%	100%	5%	67%	5%	100%	12%
0%	0%	100%	2%	100%	7%	100%	5%	67%	5%	100%	12%
0%	0%	100%	2%	100%	7%	100%	5%	67%	5%	100%	11%
0%	0%	100%	2%	100%	7%	100%	4%	100%	7%	100%	11%
0%	0%	100%	2%	100%	7%	100%	4%	100%	7%	100%	11%
0%	0%	100%	2%	100%	6%	100%	4%	100%	6%	100%	11%
0%	0%	100%	2%	100%	6%	100%	4%	100%	6%	100%	10%
0%	0%	100%	2%	100%	6%	100%	4%	100%	6%	100%	10%
0%	0%	100%	2%	100%	6%	100%	4%	100%	6%	100%	10%

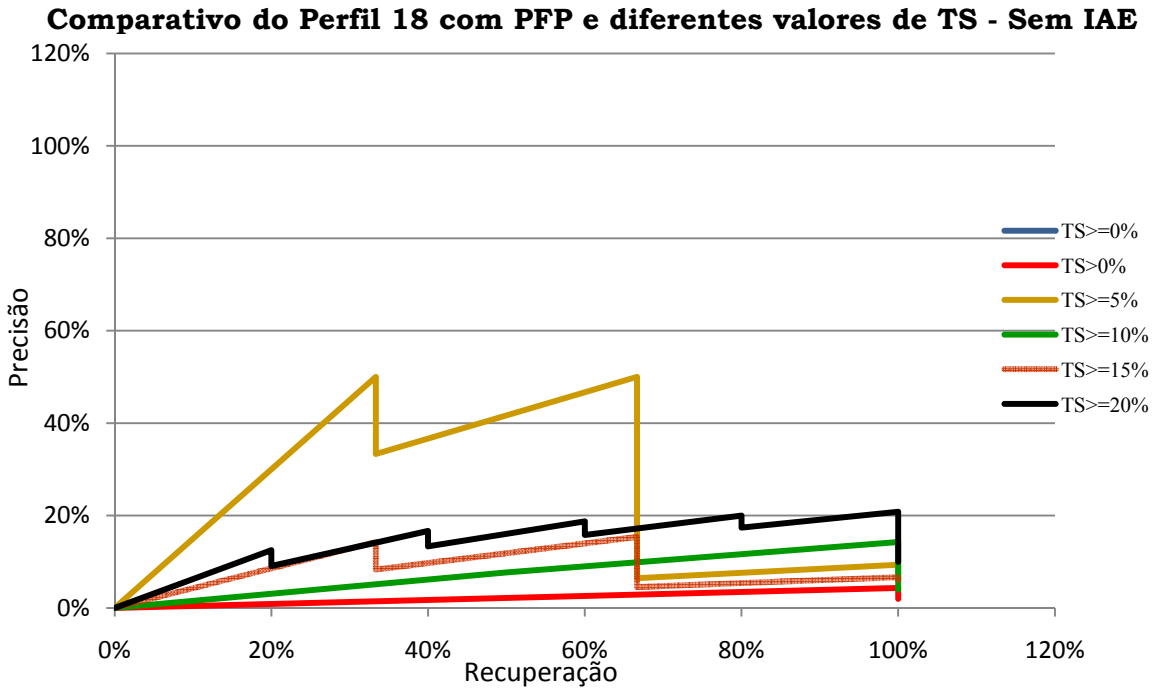


Figura 4.26 – Perfil 18: Gráfico Comparativo de Precisão X Recuperação da classificação (Sem IAE) das diferentes consultas.

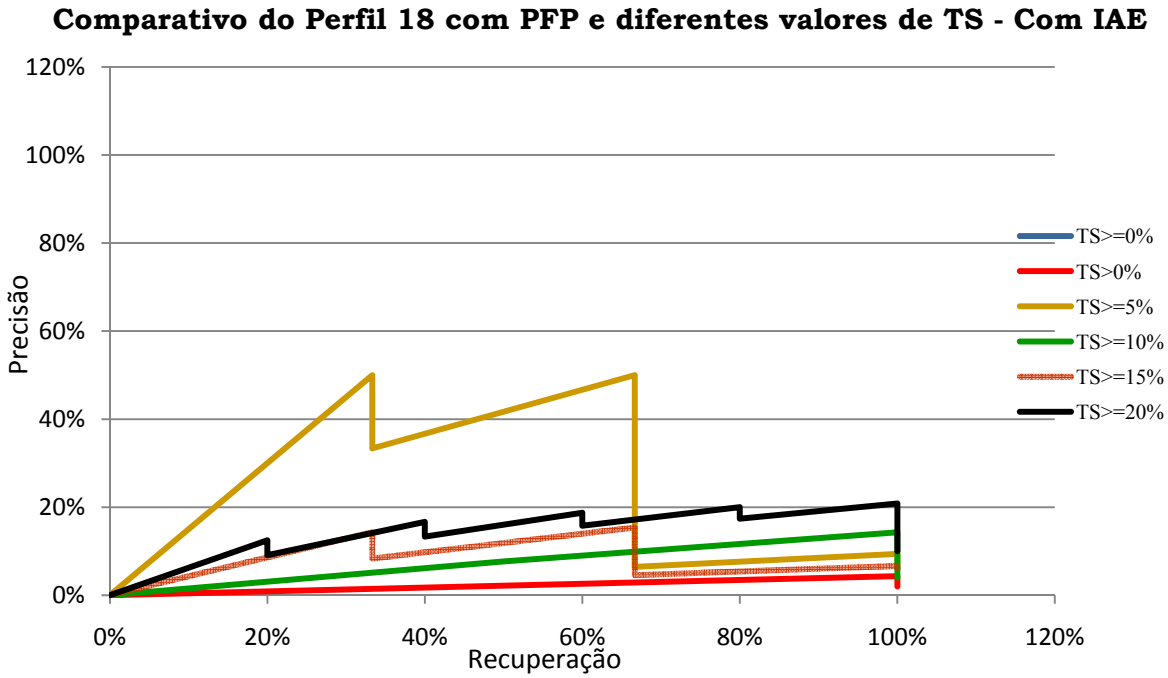


Figura 4.27 – Perfil 18: Gráfico Comparativo de Precisão X Recuperação da classificação (Com IAE) das diferentes consultas.

Note que não houve diferenças entre o uso do IAE e sem o uso do mesmo. Indica que os termos escolhidos por PFP não foram adequados para este assunto.

4.11 Validação dos Resultados.

Uma análise dos resultados levando-se em conta as dez primeiras publicações das listas de relevâncias para as Taxas de Similaridade de maior que 0% (zero por cento) e maior ou igual a 20% (vinte por cento), foi feita por 3 (três) especialistas da área de Sistemas Computacionais, sendo eles:

1. Carlos Alberto Alves Lemos – D.Sc;
2. Giancarlo Cordeiro da Costa – D.Sc;
3. José Luiz dos Anjos Rosa – D.Sc.

O laudo da análise da ferramenta da pesquisa é apresentado no apêndice B. A análise foi feita utilizando-se o perfil 12, encontrado no item referente ao assunto *Computer Science*, levando-se em conta dois conjuntos de amostras:

1. Amostra 1: As 10 (dez) primeiras publicações com uso de IAE e TS > 0% (zero por cento);
2. Amostra 2: As 10 (dez) primeiras publicações com uso de IAE e TS >= 20% (vinte por cento).

O resultado foi favorável quando da utilização da ferramenta com uso da Taxa de Similaridade mais altas, no caso, o conjunto da amostra 2, com Taxa de Similaridade maior ou igual a 20% (vinte por cento).

Capítulo 5
Conclusão

5.1. Comentários

A evolução da Internet levou a um aumento exponencial da quantidade de páginas e conteúdo armazenados. A indústria do *hardware* permite que sistemas de armazenamento sejam capazes de guardar cada vez mais dados, por sua vez, a demanda por informações, através de consultas retorna uma quantidade cada vez maior de respostas. Efetuar pesquisas sobre determinados assuntos passou a ser um grande problema, já que o tempo demandado para esta tarefa é cada vez maior. Por sua vez, novas técnicas de pesquisa passaram a ser utilizadas, mas apesar de facilitar a busca por informações, obriga a uma grande perda de tempo para ser realizada. Modelos que fazem busca por palavras chaves fez com que ferramentas de busca fossem criadas e sites de busca se tornaram importantes em nosso dia-a-dia. Os sites de busca são uma grande fonte para pesquisas, mas suas buscas são feitas sobre todo tipo de páginas, não levando em conta a relevância da publicação.

A mineração de dados se consolidou como uma ferramenta importante na recuperação de informações intangíveis em bases de dados estruturadas e suas técnicas permitiram que a mineração de textos se tornasse uma realidade como uma ferramenta de recuperação de textos correlatos. O problema é que a quantidade de textos disponíveis é muito grande e torna difícil seu armazenamento e sua recuperação. A mineração de textos na Internet passa a aumentar este problema, uma vez que os textos são buscados por palavras chaves sobre todo e qualquer tipo de documento.

Uma forma de filtragem dos documentos, através de buscas em locais específicos, um modelo de armazenamento capaz de tornar mais eficiente o emprego das técnicas de mineração de textos e a apresentação de resultados mais próximo às necessidades de grupos de usuários é fundamental para a comunidade científica. Uma ferramenta assim é tanto capaz de diminuir o tempo para aquisição de conhecimento

relevante correlato a um determinado assunto, como também identificar possíveis publicações derivadas de outros trabalhos, que podem vir a caracterizar plágios.

A metodologia aqui apresentada é capaz de automatizar o processo de pesquisa sobre documentos e de ajudar uma comunidade de pesquisadores a obter publicações interessantes sobre seu domínio de conhecimento, permitindo maior eficiência na aquisição de documentos e permitindo ainda, uma possível monitoração sobre publicações extremamente próximas, ao ponto de exigir um maior aprofundamento sobre sua originalidade.

Os resultados obtidos com os perfis 1, 2 e 8 com o número de publicações sobre o assunto escolhido entre 46% e 58% foi o qual demonstrou que a metodologia de classificação obteve um bom percentual de acerto frente aos demais seis diferentes assuntos, a aplicação do IAE resultou em ganho de Precisão e de Recuperação.

O resultado obtido com o perfil 3 teve apenas 6% de documentos sobre o assunto escolhido, o que reflete a necessidade de melhor ajuste no número de palavras relevantes a ser levado em conta por publicação. Uma melhor forma de definir este valor se torna necessária para aumentar este percentual, inclusive devendo ser ajustado em função do domínio do assunto. A aplicação do IAE permitiu que a Precisão fosse aumentada, mas reduziu um pouco a Recuperação.

O resultado obtido com o perfil 4 que não teve documentos sobre o assunto escolhido encontrados, está ligado diretamente a necessidade de melhor ajuste no número de palavras relevantes a ser levado em conta por publicação. Neste caso, não foi possível fazer a comparação utilizando o IAE.

O resultado obtido com o perfil 5 teve apenas 8% de documentos sobre o assunto escolhido, reflete também a necessidade de melhor ajuste no número de palavras relevantes a ser levado em conta por publicação. Já quanto a precisão, pode-se concluir que a perda obtida está relacionada a uma escolha ruim das palavras chaves, que foram definidas por observações sobre alguns títulos de publicações do mesmo assunto. A definição dos pesos também poderá receber uma melhor metodologia para definição.

Os resultados obtidos com os perfis 6 e 7, que tiveram 26% foi demonstrou que a metodologia de classificação obteve um bom percentual de acerto frente aos demais seis diferentes assuntos, a aplicação do IAE resultou em perda de Precisão, indicando que as definições das palavras chaves não foram adequadas e que esta metodologia necessita de melhoria. A tabela 5.1 apresenta o aproveitamento com o uso do IAE por perfil.

Tabela 5.1 – Aproveitamento da metodologia por perfil.

Perfil	**Documentos Recuperados sobre o assunto**	**Aproveitamento**	**Precisão**	**Recuperação**
1	29	58%	Aumentou	Aumentou
2	29	58%	Aumentou	Aumentou
3	3	6%	Aumentou	Diminuiu
4	0	0%	Indefinida	Indefinida
5	4	8%	Diminuiu	Estável
6	13	26%	Diminuiu	Estável
7	13	26%	Diminuiu	Estável
8	23	46%	Aumentou	Estável

Os resultados das pesquisas obtidos a partir dos perfis definidos neste trabalho revelaram que documentos classificados através do emprego da técnica de *KNN* é capaz de retornar uma lista onde trabalhos muito semelhantes que pode apresentar dúvidas sobre sua originalidade. O emprego do Índice de Aproximação Estatística (IAE) permite o aumento na Precisão e na Recuperação de documentos correlatos a um determinado domínio de assunto, diminuindo o tempo despendido com pesquisas em sites de busca e sem a leitura de grande quantidade de textos. Os resultados foram melhores quanto maior foi o aproveitamento de documentos recuperados sobre o assunto. Apesar de algumas pesquisas retornarem excelentes resultados, outros perfis mostraram que as formas de definição de número de palavras relevantes, palavras chaves por assunto e dos pesos referentes às palavras chaves podem ser melhoradas. Na próxima seção são apresentadas algumas idéias para criação de metodologias que podem ser eficazes para a resolução dos problemas apontados.

Após a aplicação da Taxa de Similaridade (TS), o problema de decorrido pela arbitragem dos valores do número mínimo de palavras relevantes foi resolvido, além de dispensar horas de reprocessamentos em ensaios de estes.

O uso da técnica de Palavras Freqüentes e Preditivas (PFP) resolveu o problema de definição dos termos relevantes por assunto e seus respectivos pesos, mas em alguns casos não surtiu o efeito esperado já que o Índice de Aproximação Estatística (IAE) obteve como resultados valores 0 (zero).

O Índice de Aproximação Estatística se mostrou eficaz quando os termos mais relevantes foram bem escolhidos, mas de qualquer forma, não atrapalha os resultados obtidos pelo seu emprego, já que não alteram os resultados finais das listas sem o seu emprego.

5.2. Trabalhos Futuros

As principais limitações encontradas foram empregadas para definição das palavras chaves e seus respectivos pesos. De forma a contribuir com trabalhos futuros que venham a se basear pelo trabalho ora apresentado, sugerem-se os seguintes itens a seguir:

- Definir uma nova metodologia para definição do número de palavras mínimas a serem utilizadas por assunto. Uma opção seria o uso de algoritmo genético (AG) para a definição destes valores. Algoritmos genéticos segundo Holland and Goldberg [40], "são procedimentos de busca probabilísticos definidos para trabalhar onde grandes espaços de busca, envolvendo estados que podem ser representados por textos.", partindo de uma população inicial para gerar um novo conjunto de indivíduos que corresponda ao ponto máximo ou mínimo da função objetivo;

- Definir uma nova metodologia para definição dos termos relevantes a serem utilizadas em cada assunto para o IAE. Esta metodologia poderia se utilizar de Ontologias para esta definição. Uma das soluções para inclusão e reconhecimento da semântica de termos é o uso de ontologias, Fonseca

e Egenhofer [41]. A semântica, ou seja, o significado dos conceitos empregados nos diversos domínios do conhecimento é capturado por meio de ontologias. Para a comunidade de Inteligência Artificial, ontologias são teorias que especificam um vocabulário relativo à um certo domínio, vocabulário este que define entidades, classes, propriedades, predicados e funções e as relações entre estes componentes [42]. Elas também são descritas como coleções estruturadas de termos, precisamente descritos e inter-relacionados entre si, de acordo com o entendimento de uma comunidade de especialistas no domínio. Com o uso de ontologias, é possível representar informações que refletem um entendimento semântico de diversas situações do mundo real.

- Utilizar um dicionário de sinônimos na preparação dos dados para aumentar a possibilidade de descoberta de publicações não originais, ou informações inesperadas. Segundo BASTOS [19], uma informação é considerada inesperada, "... quando ela é relevante, porém desconhecida do usuário, ou contradiz a opinião existente ou as expectativas do usuário.

Referências Bibliográficas

[1] MARTINS, Wilson. *A palavra escrita. História do livro, da imprensa e da biblioteca*. São Paulo: Ática, 2002.

[2] GIOVANNINI, Giovanni. *Evolução na Comunicação. Do sílex ao silício*. Rio de Janeiro: Nova Fronteira, 1987.

[3] ANUNCIAÇÂO, H.S. *Linux para redes brasileiras*. São Paulo, Érica, 1997.

[4] MATTELART, A. M. *História das teorias da comunicação.* P. 65-66. São Paulo: Edições Loyola, 1999.

[5] Google. Disponível em HTTP://www.google.com.br.

[6] AGICHTEIN E., ZHENG Z..: *Identifyind "Best Bet"Web Search Results by Mining Past User Behavior. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM Press New York, NY, USA, 2006.

[7] JOACHIMS T., GRANKA, L., PANG, B., HEMBROOKE H. *and* Gay G.,: *Accurately Interpreting Clickthrough Data as Implicit Feedback*. In the Proceedings of SIGIR, 2005.

[8] VIEIRA, Pedro Antônio. *A História da Automatização do Trabalho.* Disponível em: <HTTP://orbita.starmedia.com/~novosdebates/textos/t10.htm>. Acesso em 25 out. 2005.

[9] TAN A.H.: *Text Mining the State of the Art and the Challenges.* In Proceedings of the PAKDD 1999. Workshop on Knowledge Disocovery from Advanced Databases, pages 65–70, Beijing, China, April 1999.

[10] ROSA, J. L. A.: *Classificação de Dados Através da Otimização do Método KNN-Fuzzy em Ambiente de Computação Paralela.* Rio de Janeiro, 2003.

[11] REZENDE, Solange Oliveira. *Org. Sistemas Inteligentes: fundamentos e aplicações.* Pág. 337. São Paulo: Manole, 2003.

[12] SPSS - Disponível em: http://www.spss.com.br/tecnologias/text_mining.htm.

[13] SAS Text Miner - Disponível em: (http://www.sas.com/technologies/ analytics/datamining/ textminer/#section=1)

[14] STATISTICA - Disponível em: http://www.statsoftiberica.com/pt/solucoes/ produtos/pacstat_datatextaminer.html

[15] ORACLE - Disponível em: http://download-west.oracle.com/docs/cd/ B12037_01/datamine.101/b10698/8text.htm#1006650

[16] SQL Server - Disponível em: http://www.microsoft.com/Sqlserver/ 2005/en/us/Compare-Features.aspx

[17] PINHEIRO, Marcello Sandi. Uma Abordagem Usando Sintagmas Nominais Como Descritores no Processo de Mineração de Opiniões. Rio de Janeiro: UFRJ/COPPE, 2009.

[18] PINHEIRO, Marcello Sandi. Uma Abordagem Usando Sintagmas Nominais Como Descritores no Processo de Mineração de Opiniões. Rio de Janeiro: UFRJ/COPPE, 2009. Op cit 23.

[19] BASTOS, Valéria M. Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa. Rio de Janeiro: UFRJ/COPPE, 2006.

[20] CHAVES, M. S.: *Um estudo e apreciação sobre algoritmos de stemming para a língua portuguesa.* IX Jornadas Iberoamericanas de Informática. Cartagena de Indias - Colômbia, 11-15 agosto de 2003.

[21] FRAKES, W.B., BAEAZA-YATES, R.: *Readings in Information Retrieval: Data Structured Algorithmos.* Ed. Upper Saddle River, NJ: Prentice Hall, 1992.

[22] WIVES, L.K.: *Indexação de Documento.,s Textuais.* Disponível em: http://www.inf.ufrgs.br/~wives/publicacoes/IDT.pdf.

[23] LOPES, M. C. S.: *Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português.* Rio de Janeiro, 2004.

[24] SPARK-JONES, K., WILLET, P.: Readings in Information Retrieval. San Francisco: Morgan Kaufmann, 1997.

[25] BAEZA-YATES, R.: *Modern Information Retrieval.* New York, N.Y.: Addison-Wesley, 1999.

[26] MARTHA, A. S. BARRA, P.S.C., CAMPOS, C. J. R.: *Recuperação de Informações em Textos Livres de Prontuários do Paciente.* Disponível em: http://www.sbis.org.br/sbis/arquivos/636.pdf.

[27] LOVINS J. B.: *Development of a Stemming Algorithm. Maechanical Translation and Computacional Linguistics*, 11, 22-31, 1968.

[28] PORTER, M. F.: *Stemming Algorithm Parper*. Computer Laboratory Canbridge England, 1979.

[29] PORTER, M. F.: *The Porter Stemming Algorithm.* Disponível em: http://www.tartarus.org/~martim/PorterStemmer/index.html.

[30] HARMAN, D.: How *Effective is Suffixing?*. Journal of the American Society for Information Science. 42(1): 7-15, 1991.

[31] PORTER, M. F.: *An Algorithm for Suffix Stripping*. Program, 14(3), 130-137, 1980.

[32] SILVA, Jaqueline Uber. *Text Mining com uma aplicação na validação dos registros de ocorrências policiais na região da grande Florianópolis*. MSc. 2005, Universidade Federal de Santa Catarina, 123 p.

[33] FERNEDA, Edberto. Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. São Paulo: USP, 2003.

[34] SILVA, Jaqueline Uber. *Text Mining com uma aplicação na validação dos registros de ocorrências policiais na região da grande Florianópolis*. MSc. 2005, Universidade Federal de Santa Catarina. Op cit 34-35.

[35] ZUCHINI, Márcio Henrique. *Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informação.* Dissertação de mestrado. Campinas, SP: [s.n], 2003.

[36] MATSUNAGA, Liliam A., Uma Metodologia de Categorização Automática de Textos para a Distribuição dos Projetos de Lei às Comissões Permanentes da Câmara Legislativa do Distrito Federal. Rio de Janeiro: UFRJ/COPPE, 2007.

[37] POPESCUL, a.; UNGAR L.H., *Automatic Labeling of Document Clusters*. Disponível em: http:// www.cis.upenn.edu/~popescul/Publications/ popescul00labeling.pdf

[38] LIU, Bing. *Web data mining: exploring hyperlinks, contents, and usage data (data-centric systems and applications).* New York: Springer-Verlag, 2006.

[39] EBECKEN, N.F.F.; Lopes, M.C.S.; COSTA, M.C.A. *Sistemas Inteligentes Fundamentos e Aplicações* / organização, Solange Oliveira Rezende. – Barueri, SP: Manole, 2003.

[40] HOLLAND,JH.; GOLDBERG, D.E., *Genetic Algorithms and Machine Learning Springer Netherlands*, cit. 95-99, 2005.

[41] FONSECA, F.; EGENHOFER, M.; BORGES, K.. *Ontologias e Interoperabilidade Semântica entre SIGs.* Disponível em: http://www.geoinfo.info/geoinfo2000/papers/011.pdf

[42] GRUBER, T.R.. *What is an ontology?* Disponível em: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html, 1992.

Apêndice A

Tabela de *stoplist* com palavras universais para a língua inglesa utilizada neste trabalho de pesquisa:

about	add	Ago	after	all
also	an	And	another	any
are	as	at	be	because
been	before	being	between	big
both	but	by	came	can
come	could	did	do	does
due	each	else	end	far
few	for	from	get	got
had	has	have	he	her
here	him	himself	his	how
if	in	into	is	it
its	just	let	lie	like
low	make	many	me	might
more	most	much	must	my
never	no	nor	not	now
of	off	old	on	only
or	other	our	out	over
per	pre	put	re	said
same	see	she	should	since
so	some	still	such	take
than	that	the	their	them
then	there	these	they	this
those	through	to	too	under
up	use	very	via	want
was	way	we	well	were
what	when	where	which	while
who	will	with	would	yes
yet	you	your		

Apêndice B

A seguir o laudo de análise da ferramenta de pesquisa.

Rio, 21 de dezembro de 2009.

A ferramenta "Pesquisa, Busca e Armazenamento de Dados Semi-Estruturados na Web" é uma ferramenta que busca, por meio de mineração de textos e cálculo de distância, artigos com alto grau de similaridade, baseado no contexto de um documento base (denominado "perfil").

Para a análise da ferramenta, foram aplicados testes com duas amostras, utilizando-se, como parâmetros de configuração, duas taxas de similaridade: a primeira, para a **Amostra 1**, maior que zero (TS > 0%), e a segunda, para a **Amostra 2**, maior ou igual a 20% (TS ≥ 20%).

A **Amostra 1** apresentou baixo rendimento de recuperação de documentos correlatos, obtendo um conjunto com baixo grau de relevância entre eles.

Com relação ao grau de relevância dos documentos recuperados, a **Amostra 2** apresentou um melhor rendimento, o que era se de esperar, pois, neste caso, o filtro imposto exigiu o aumento da correlação entre os itens ora em pesquisa.

Os resultados desses dois testes, agregados aos resultados obtidos durante os trabalhos de doutoramento, permite à equipe de especialistas, cujas assinaturas seguem abaixo, emitir laudo favorável à ferramenta quanto à sua capacidade de efetuar mineração de textos de forma satisfatória, com a respectiva obtenção de documentos correlatos baseando-se em um documento de perfil.

Carlos Alberto Alves Lemos, D.Sc.

Giancarlo Cordeiro da Costa, D.Sc.

José Luiz dos Anjos Rosa, D. Sc.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo