



SISTEMA DE INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM CONSTRUÇÃO BASEADO EM ONTOLOGIA

AMORIM, Sergio R. Leusin (1); CHERIAF, Malik (2)

(1) UFF - Universidade Federal Fluminense, Rua Passo da Pátria, 156, Bl. D s.541, Niterói, RJ
e-mail: leusin@poscivil.uff.br

(2) Universidade Federal de Santa Catarina – UFSC -CTC – ECV, Trindade - Florianópolis - Santa
Catarina - Brasil - CEP 88040-900 e-mail: malik@infohab.org.br

RESUMO

Este trabalho descreve o desenvolvimento de um sistema de indexação e recuperação de informação na área de AEC- Arquitetura, Engenharia e construção, visando a melhorar a precisão e revocação na busca de documentos. Baseando-se na classificação de objetos do universo construído desenvolvido no âmbito do projeto CDCON e na ontologia destes objetos proposta pelo projeto ONTOARQ, o sistema promove uma indexação automática da base de documentos. Após a aplicação de ferramentas específicas é feita uma análise de relevância tomando-se por base os termos classificados pela estrutura CDCON e interrelacionados na base do ONTOARQ. As associações ontológicas desta base permitem estender a busca aos termos associados, melhorando o desempenho do sistema nos aspectos de precisão e revocação na busca de documentos. O sistema foi implantado em protótipo, utilizando-se a base de documentos PDFs existente no INFOHAB, Centro de Referência e Informação em Tecnologia do Habitat. Os resultados indicam que o sistema proposto melhora a eficácia da recuperação de informação no âmbito do banco de dados do INFOHAB, traduzida em melhora quantitativa de 35% na precisão média.

ABSTRACT

This work describes the development of an indexation and recovery of information system in the AEC- area Architecture, Engineering and construction, aiming to improve the precision and revocation in the document search. Based on construction's object classification developed in the scope of project CDCON and in the ontology proposed in ONTOARQ project, the system promotes an automatic indexation of the INFOHAB documents database. After the application of tools to arrive to normalize terms, it is made a relevance analysis comparing the terms classified in CDCON structure with its associations in ONTOARQ database. This ontology base allows to extend to the search to the terms associates, improving the document search performance regarding precision and revocation aspects. The system was implanted as a prototype, having used the existing PDFs document in the INFOHAB database. The results indicate that it improves the effectiveness of information recovery, translated in a quantitative improvement of 35% of mean accuracy.

Palavras-chave: recuperação de informação, classificação de documentos, ontologia, construção.

1. INTRODUÇÃO

Sistemas de recuperação de informações – SRI ou, em Inglês, IRS (Information Retrieval System) objetivam o armazenamento, recuperação e gerenciamento de informações. Informação, neste contexto, pode ser composta de textos (incluindo o formato numérico e datas), imagens, áudio, vídeo e outros objetos multimídia. Sistemas baseados em reconhecimento de imagens são complexos, estando ainda em fase de uso restrito. Já os sistemas de áudio muitas vezes baseiam-se na sua transformação em texto, à exceção daqueles voltados ao mercado musical. Em ambos os casos é comum a utilização de metadados para classificação dos arquivos de áudio ou imagem, sendo a recuperação baseada nestes campos apenas. Dada a relevância de textos, os maiores esforços têm sido direcionados a este tipo de formato.

Embora a tecnologia básica de recuperação de informações arquivadas em bases de texto extensas seja bastante comum, ela ainda apresenta resultados aquém dos desejados. Nos sistemas genéricos, como os buscadores comuns na INTERNET, são comuns respostas sem relevância ao interesse real do usuário, pois não consideram o contexto específico. Trata-se de um problema complexo e objeto de diversos estudos e propostas, onde se pode destacar, vinculados à textos de AEC, os trabalhos de NASCIMENTO (2004) e CALDAS (2002), entre outros

Mesmo em bases focadas em uma determinada área de domínio existem diferenças de contexto de cada busca que levam a respostas muito extensas e fora do interesse do pesquisador. Sistemas que apresentem melhor desempenho facilitam as buscas e colaboram para melhores ferramentas de gestão de conhecimento, recursos humanos e outras. O desempenho é usualmente relacionado à precisão e à revocação, sendo a primeira a expressão da quantidade de repostas relevantes na lista de resultados e a segunda a relação entre a quantidade de respostas relevantes apresentadas comparadas a todas as relevantes existentes na base de dados.

2. SISTEMAS DE RECUPERAÇÃO TEXTUAL

Um sistema de recuperação de informações textuais é um sistema desenvolvido para indexar e recuperar documentos do tipo textual, ou seja, documentos cujas informações estão descritas através da linguagem natural. São sistemas que tratam basicamente informações do tipo texto (ASCII), mas, que através de filtros adequados podem analisar outros formatos que contenham textos, figuras, tabelas e imagens, mas que possuam um aspecto de documento textual, como o PDF, o PS ou DOC). Note-se que a parte de imagem destes documentos será descartada pelos filtros, embora ferramentas mais sofisticadas possam identificar legendas e, por associação, recuperar as imagens.

Para abreviar o tempo de processamento estes sistemas efetuam uma catalogação e classificação prévia dos documentos, atualizada a cada nova inserção ou em intervalos pré-determinados. Esta classificação gera um índice que reflete a relevância do documento no contexto da busca. Este processo é descrito por LANCASTER (1985 1987).

O vocabulário empregado em um sistema de recuperação deve ser um vocabulário controlado, caracterizado por um conjunto limitado de termos, os quais se encontram organizados em alguma forma de estrutura que permita controlar sinônimos e remissivos que indiquem relações entre os termos (LANCASTER, 1985). Deste modo, um sistema de recuperação possui duas bases de dados distintas: uma armazena o conjunto de documentos, dos quais se deseja obter informações, e a outra contém as entradas que representam os documentos do sistema. Estas entradas são os descritores obtidos no processo de indexação, podendo ser considerado como um índice da outra base de dados (LANCASTER, 1985).

Esta estrutura guarda em si dois problemas: a indexação é trabalhosa e se realizada manualmente muitas vezes depende da subjetividade do classificador. Se automática, necessita de um vocabulário de referência que traduza efetivamente o contexto da busca.

Mesmo a indexação automática demanda tempo e capacidade de processamento, daí a importância de uma estrutura lógica para facilitar generalizações e melhorar o desempenho. A forma mais usual é a construção de um tesauro específico para a área de domínio ensejada. Neste caso, optamos pelo tesauro desenvolvido no projeto CDON (AMORIM, 2006)

A partir desta base, o processo de indexação automática compara e identifica os termos relevantes (descritores) nos documentos de uma coleção e os insere em uma estrutura de índice. As fases normalmente encontradas nesse processo são a identificação de termos (simples ou compostos), a remoção de *stopwords* (palavras irrelevantes), a normalização morfológica (*stemming*) e a seleção de termos (Krug 2004).

Este processo é bem descrito nesta literatura técnica, assim como a discussão sobre a ponderação que cada documento deve receber ao ser inserido no índice. Existem três formas para esta análise: a Frequência absoluta de termos (*term frequency* ou *absolute frequency*) Frequência relativa de termos (*relative frequency*), Frequência inversa de documentos (*Inverse document frequency*) A análise de relevância é realizada através da função de *similaridade*, mas esta comparação entre termos consultados e documentos em geral traz documentos irrelevantes. Para melhorar estes resultados foram propostos modelos conceituais de recuperação, posteriormente adaptados às ferramentas de busca: o modelo booleano, o espaço-vetorial, o probabilístico, o de busca direta, o de aglomerados (clusters) e o contextual ou conceitual.

Ainda assim, o desempenho dos sistemas de recuperação de informação depende primordialmente da organização e estrutura da base de dados de referência. Se ela refletir melhor o contexto do universo pesquisado, ela resulta em melhor precisão e revocação do sistema.

3. A BASE RELACIONAL DO ONTOARQ

O projeto CDCON (AMORIM, 2006) propôs uma estrutura facetada para a descrição dos objetos construídos. Utilizando estes conceitos teóricos foi desenvolvido um sistema – o ONTOARQ, para o desenvolvimento e gerenciamento das relações associativas entre conceitos e seus termos descritores. Este sistema, disponível em <http://www.moleque.com.br/TesaurusUff/>, vem permitindo o aperfeiçoamento de uma ontologia em AEC- Arquitetura, Engenharia e Construção. Através dele foi possível identificar regras de relações associativas, que facilitam não só o próprio desenvolvimento da ontologia como podem servir de base para buscas lógicas, mais eficientes.

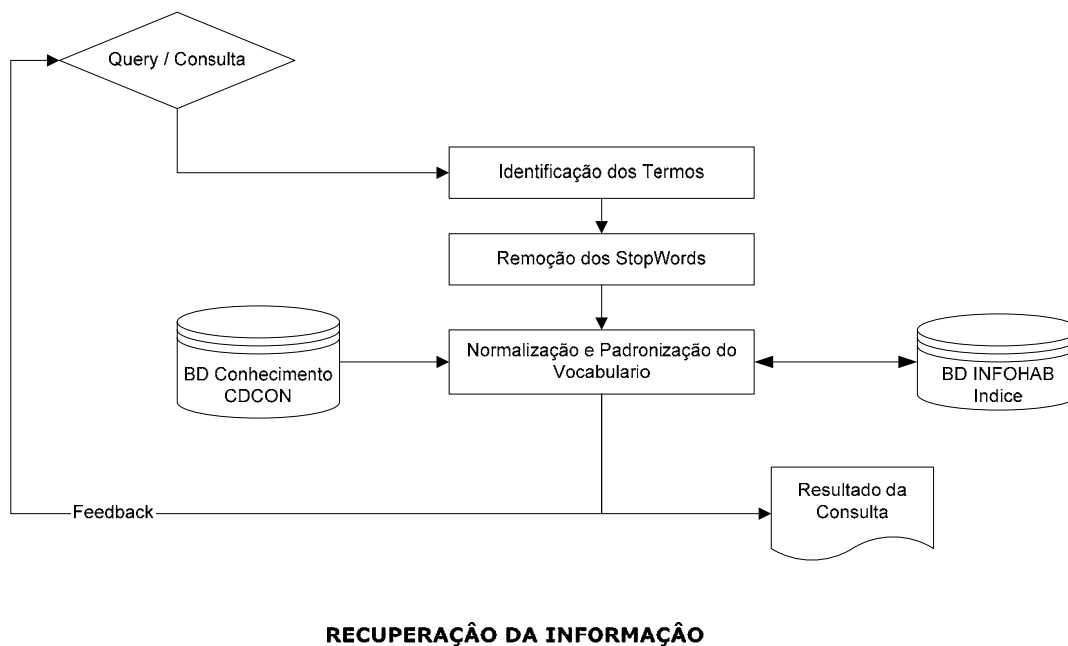
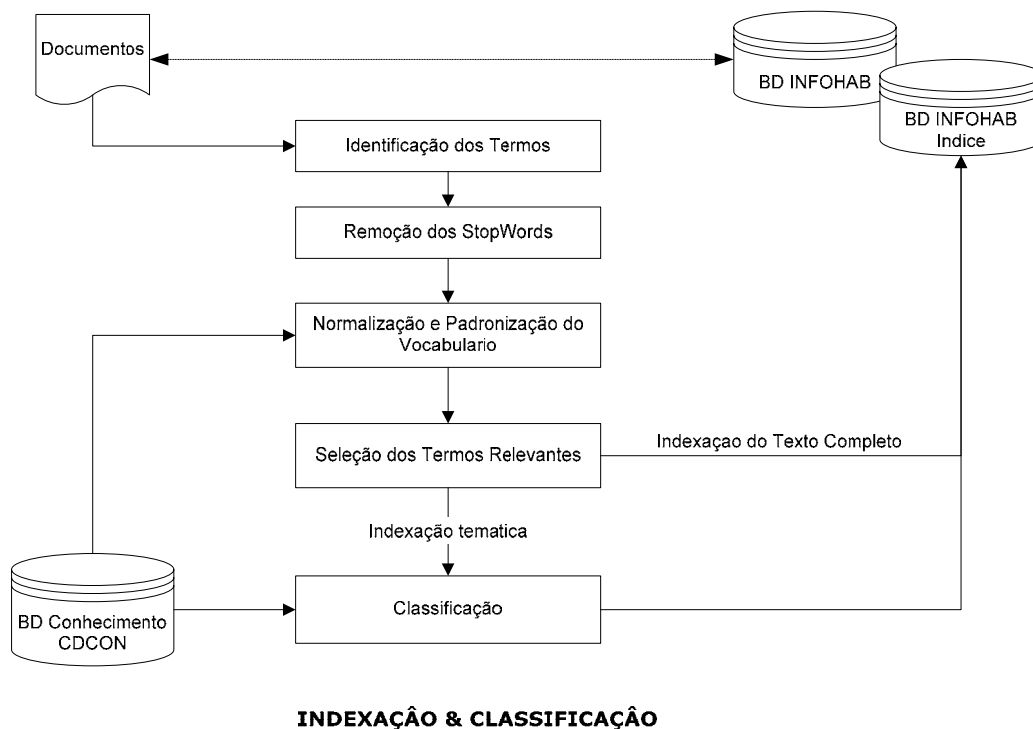
Esta base de dados relacional, estruturada em facetas foi utilizada como referência para o processo de indexação, resultando em melhoria de desempenho, como descrito adiante.

4. METODOLOGIA

A proposta metodológica deste projeto visa o desenvolvimento de um sistema de recuperação da informação no INFOHAB baseado em sistema de classificação das informações o CDCON, único sistema de classificação para a construção civil no BRASIL. O objetivo é implementar de um lado um sistema de indexação automático dos documentos catalogados INFOHAB e de outro lado melhorar o mecanismo de busca atual do INFOHAB.

A metodologia de indexação escolhida é aquela mostrada na **Figura 1** que consiste na identificação de termos (simples ou compostos), a remoção de *stopwords* (palavras irrelevantes), a normalização morfológica (lemetização e *stemming*) e a seleção de termos baseada num lado no sistema de classificação do CDCON e de outro lado pela medição da importância de cada palavra, identificando seu “peso” ou “força” de representatividade (*term weight*) considerando-se a frequência das palavras nos documentos. Também no sistema de recuperação da informação, foi utilizado no pré-processamento da consulta a mesma metodologia de indexação e a busca será realizada nas tabelas dos índices.

Figura 1 Esquema do SRI



5. RESULTADOS

O desenvolvimento foi realizado em três etapas:

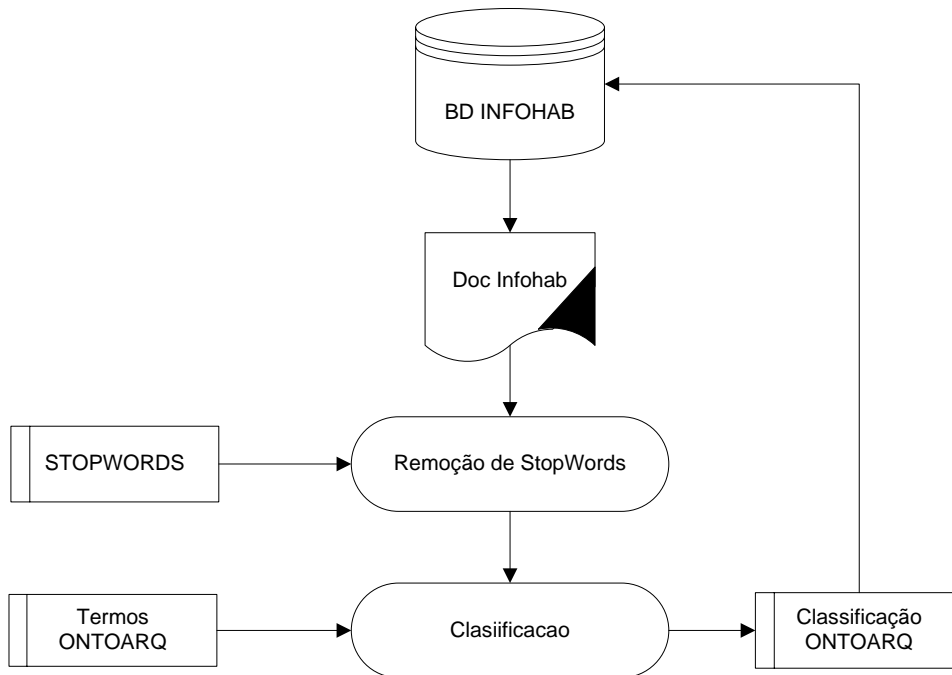
- Integração do sistema de classificação de CDCON, através da base de dados extraída do ONTOARQ, no INFOHAB e Indexação da Base de dados e os documentos atuais do INFOHAB para gerar os Índices
- Extração dos termos do banco de dados do INFOHAB com avaliação da relevância destes e classificação dos documentos do INFOHAB usando estes termos.

- Implementação do Sistema de Recuperação da Informação

5.1. Integração do Sistema de Classificação de CDCON - ONTOARQ no INFOHAB e Indexação da Base de dados

Nesta etapa (*Figura 2*) foi realizada a importação da lista dos termos com as facetas correspondentes extraídas do ONTOARQ. Após a extração dos termos e remoção dos stopwords (anexo 02), para cada documento do banco de dados de INFOHAB, no texto resultante são localizados e identificados os termos possíveis do sistema ONTOARQ para constituir um novo index dentro do banco de dados do INFOHAB.

Figura 2 Integração do sistema de classificação de CDCON

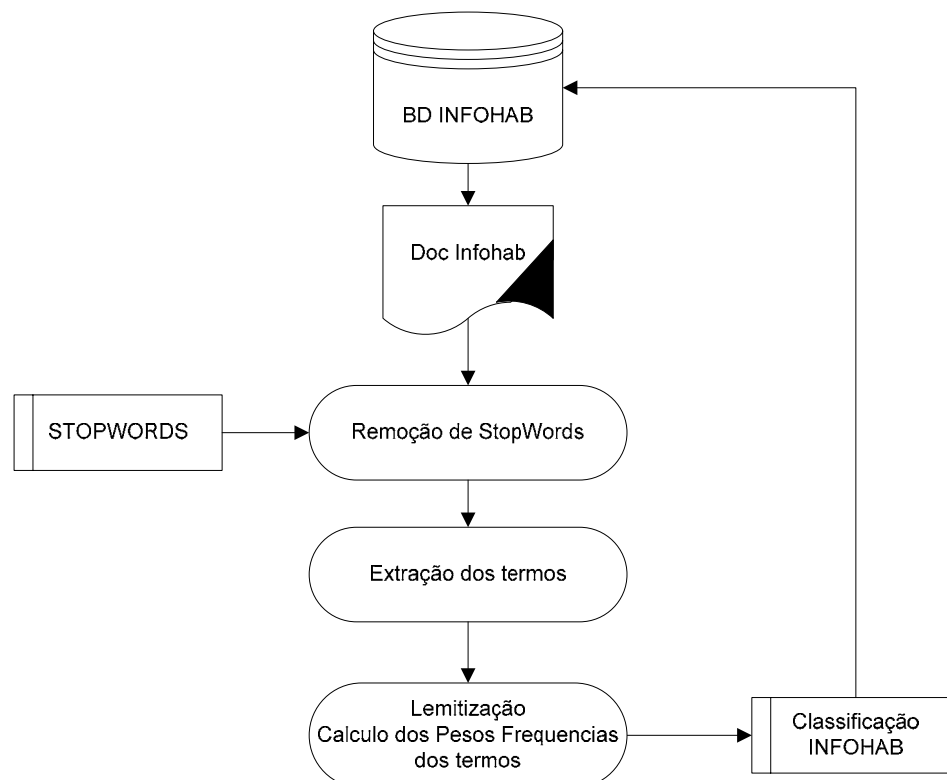


5.2. Extração dos Termos do Banco de dados INFOHAB

Nesta etapa (*Figura 3*), após a extração dos termos e remoção dos *stopwords*, para cada documento do banco de dados de INFOHAB, no texto resultante cada palavra é localizada e é identificada pelo numero de ocorrência e frequência relativa e absoluta. Neste processo é gerado um novo índice de classificação a partir das palavras extraídas que registrado dentro do banco de dados de INFOHAB

Foram identificadas 813.356 palavras que após a redução e remoção dos *stopwords*, foram reduzidas a 59347. Considerando uma frequência inversa próxima de zero este numero pode ser reduzido para 49367 palavras correspondente a uma redução do índice de 94%. Ressalta-se que na maioria dos sistemas de busca é esta a metodologia de extração empregada. Usando o processo de lemetização este número de termos fica finalmente reduzido a 12440.

Figura 3 Extração dos Termos do Banco de dados INFOHAB

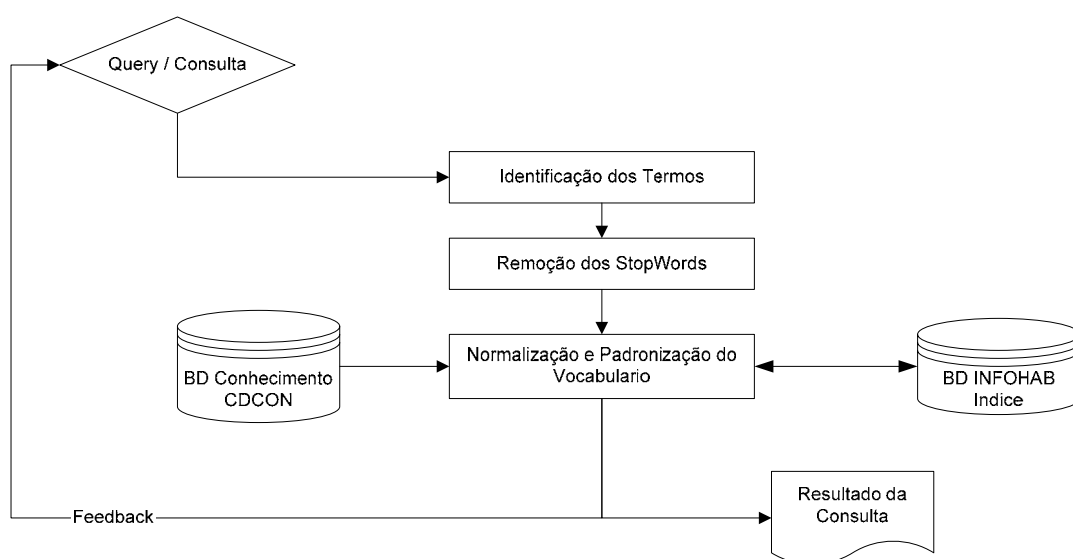


5.3. Implementação do Sistema de Recuperação das Informações

A proposta metodológica para recuperação de informações é baseada em um sistema de classificação das informações da construção civil (*Figura 4*). Estas informações abrangem todos os processos do ciclo de vida de uma construção com classificações relacionadas aos processos, fases, materiais, produtos e espaços. O objetivo desta ferramenta é melhorar o resultado dos mecanismos de busca convencionais. Para isso são propostas a expansão dos argumentos de busca e a ordenação da apresentação dos resultados por ordem de relevância dos documentos recuperados (ranking) através da adoção de pesos. Estes documentos serão apresentados através de uma lista com nomes e descrições de documentos que apontam para o documento original (arquivo eletrônico) no banco de dados do sistema. A expansão dos argumentos de busca será por meio de operações de expansão de termos por relacionamento de equivalência e por relacionamento hierárquico. Estas operações do mecanismo de busca levam em conta as especificidades da indústria da construção civil o que tornando estes mecanismos mais eficazes do que aqueles obtidos com técnicas convencionais genéricas.

O funcionamento do sistema consiste em processar uma consulta em um mecanismo de busca para que um usuário obtenha uma resposta à sua necessidade de informação, obtendo documentos armazenados. A consulta é feita por meio de um formulário onde o usuário pode usar linguagem natural ou palavras-chave. A busca não é feita no conteúdo dos arquivos no banco de dados do sistema, mas sim nos descritores dos documentos. Os descritores são informações sobre o conteúdo dos arquivos de documentos. A eficácia do sistema de recuperação de informação depende em grande medida da qualidade dos argumentos de busca especificados pelo usuário. Idealmente devem ser específicos e em boa quantidade. O pequeno número de argumentos de busca digitados pelo usuário talvez seja um dos maiores problemas para obter-se uma maior eficácia do sistema.

Figura 4 Implementação do sistema de recuperação das informações



Os termos dos argumentos das consultas precisam ser preparados antes de serem utilizados pelos algoritmos da ferramenta de busca. Os primeiros passos deste pré-processamento são as operações sobre texto. Inicialmente, o texto da consulta é carregado para a ferramenta e separado em termos para ser montada uma lista. Em seguida, os caracteres retirados dos argumentos da consulta passam pelos processos de Análise Léxica, eliminação de *Stopwords*, Lematização (*Stemming*) resultando na criação de uma lista de termos.

6. RESULTADOS E CONCLUSÕES

O

Quadro 1 compara o sistema de busca atual do INFOHAB com o sistema desenvolvido. Em quase todos os termos analisados o novo sistema de recuperação de informação (<http://143.107.96.102>) retorna um número de documentos relevantes superior ao sistema atual do INFOHAB (www.infohab.org.br).

Quadro 1: Comparação de resultados

Termo	Busca INFOHAB	Busca Indexada
ADITIVO SUPERPLASTIFICANTE	3	312
AGLOMERANTE	21	103
AGREGADO	266	409
ANTEPROJETO ARQUITETÔNICO	0	31
AZULEJO	11	21
CADERNO DE ENCARGOS	2	17
GIPSITA	7	15
ILUMINAÇÃO	397	249
LADRILHO	16	7
MAQUETES	4	23
TRINCO	9	9
TUBO DE PVC	48	305

Estes resultados indicam que o sistema proposto melhora a eficácia da recuperação de informação no âmbito do banco de dados do INFOHAB. Esta eficácia é traduzida quantitativamente em melhora de 35% de precisão média. O cálculo dos pesos para ordenação dos documentos e o processo de busca contextual mostrou-se importantes para a melhoria da precisão. A expansão dos termos de busca

através de generalização e especialização e o processo de lematização mostraram-se eficazes para melhorar a revocação do sistema. Para conseguir recuperar mais eficientemente documentos é necessário descrever detalhadamente os documentos armazenados nos descritores.

BIBLIOGRAFIA:

NASCIMENTO, Luis Antonio; **Proposta de um Sistema de Recuperação de Informação para Extranet de Projeto**, Dissertação, Escola Politécnica da Universidade de São Paulo, USP, SP, 2004

CALDAS Carlos H., SOIBELMAN S.M.; Lucio; HAN Jiawei; **Automated Classification of Construction Project Documents**; JOURNAL OF COMPUTING IN CIVIL ENGINEERING, Volume 16, Issue 4, pp. 231-299 ,2002

LOPES, Ilza Leite. **Search strategy in information retrieval: literature review**. *Ci. Inf.*, May/Aug. 2002, vol.31, no.2, p.60-71. ISSN 0100-1965.

BOMBASSARO JÚNIOR, A. **Construção de uma metodologia baseada em análise de conteúdo para criação de ontologias para sistemas de informações geográficas, visando maior interoperabilidade semântica**. 2003. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas (ICET), Centro Universitário FEEVALE, Novo Hamburgo.

LOH, Stanley ; CASTRO, F. M. ; OLIVEIRA, José Palazzo Moreira de ; SILVA, A. C. M. ; WIVES, Leandro Krug ; LICHTNOW, D. . **Investigação sobre a construção de ontologias a partir de textos com suporte de ferramentas automatizadas**. In: I WORKSHOP DE ONTOLOGIAS, 2002, São Leopoldo, 2002

F. Wilfred Lancaster. **Vocabulary control for information retrieval**. Information Resources Press, Arlington, VA, 1987.

LANCASTER, F. W. **Construção e uso de tesouros: curso condensado**. Brasília: IBICT, 1987. 106 p.