



INTERCOM – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação
XXV Congresso Brasileiro de Ciências da Comunicação – Salvador/BA – 1 a 5 Set 2002

Especificação de um Índice de Citação para Biblioteca Digital Auto-incremental¹

Igor Campos Leal
Universidade Federal de Minas Gerais
igorleal@dcc.ufmg.br

Marcello Peixoto Bax
Universidade Federal de Minas Gerais
bax@ufmg.br

Resumo

Acompanhando a evolução da Web, os sistemas de recuperação de informação tornam-se cada vez mais importantes no dia a dia das pessoas, já que para encontrar informações relevantes no contexto atual é necessária a utilização de ferramentas que auxiliem a recuperação de informações. Este artigo aborda a especificação de uma ferramenta denominada “Índice de Citação Autônomo”, módulo componente de uma biblioteca digital auto-incremental que é objeto de projeto na Escola de Ciência da Informação da UFMG. O principal aspecto tratado no artigo é a modelagem de dados dessa ferramenta, necessária para a finalidade de indexar as citações feitas em artigos técnicos e científicos, relacionando os artigos da coleção entre si e com as palavras chave utilizadas; seja por citação explícita entre os autores, seja por semelhança de conteúdo.

Palavras chave: bibliotecas digitais, recuperação de informação, índice de citações.

2 Introdução

Nos anos que precederam a Internet, a comunidade científica e tecnológica ressentia-se da dificuldade em ter acesso à literatura científica, seja porque o acesso era muito dispendioso, dado o elevado custo das bases de dados científicas mais importantes, seja porque as tecnologias de recuperação de informação e comunicação, na época, eram ainda rudimentares e não ajudavam muito o acesso direto a tais informações [4].

Com o avanço das Tecnologias da Informação e da Comunicação e o advento da Internet, o registro e a disseminação da informação científica e tecnológica sofreram grandes impactos como, por exemplo, o volume de informação registrado que cresceu exponencialmente. Segundo Lawrence [1], a rede Web em 1998 atingia 800 milhões de páginas. Além disso, o usuário passou a ter acesso direto à informação sem a intermediação de um profissional da informação, inclusive para os autores, que passaram a ter maior facilidade em tornar disponíveis os resultados de suas pesquisas, livrando-se, de certa maneira, da dependência dos

¹ Trabalho apresentado no XII ENDOCOM, XXV Congresso Anual em Ciência da Comunicação, Salvador/BA, 05. setembro.2002.



grandes editores de revistas e jornais científicos. Desta forma, para encontrar informações relevantes é necessária a utilização de ferramentas que auxiliem a recuperação de informações.

Esse conjunto de fatores tem levado vários grupos de pesquisa a estudar, propor e desenvolver ferramentas que auxiliem os usuários a satisfazer suas necessidades de informação. É neste contexto que se insere o presente projeto que se propõe a realizar pesquisa na área de bibliotecas digitais com vistas à especificação, o desenvolvimento e a implantação de uma ferramenta na Web que facilite ao usuário da comunidade científica e tecnológica satisfazer suas necessidades de informação de forma mais precisa.

Vale ressaltar que o trabalho aqui apresentado faz parte de um projeto maior, coordenado por dois professores da Escola de Ciência da Informação da UFMG: *Pesquisa e Desenvolvimento de Ferramentas para a Captura, Tratamento e Recuperação de Literatura Científica na Web* [4], composto por vários alunos do Departamento de Ciência da Computação e de Ciência da Informação. O Índice de Citação Autônomo é um módulo deste projeto, e cada aluno estará incumbido de desenvolver um módulo diferente, ou seja, cada módulo independe do resultado do outro e estes módulos se complementam formando um pacote completo.

O desenvolvimento da especificação da ferramenta denominada Índice de Citação Autônomo (ICA), relatado neste artigo, focalizou-se no desenvolvimento de um modelo de dados que possibilita a concepção de uma rede de citações que permita identificar e recuperar os artigos citados e citantes de um determinado documento automaticamente.

Este documento apresentará na Seção 2 a especificação detalhada do problema e na Seção 3 o objetivo do trabalho descrito neste artigo. A Seção 4 explicitará a estrutura deste trabalho, contemplando todas as etapas de vida do processo (entrada de dados – processamento dos dados – saída dos resultados). A Seção 5 descreverá como os requisitos funcionais levantados na Seção 3 serão providos, demonstrando o modelo de dados proposto. Já, na Seção 6 será discutido o assunto sobre o desempenho da modelagem proposta e na Seção 7 serão apresentados alguns trabalhos relacionados. Por fim, a Seção 8 apresentará uma conclusão deste relatório.

3 Especificação do problema

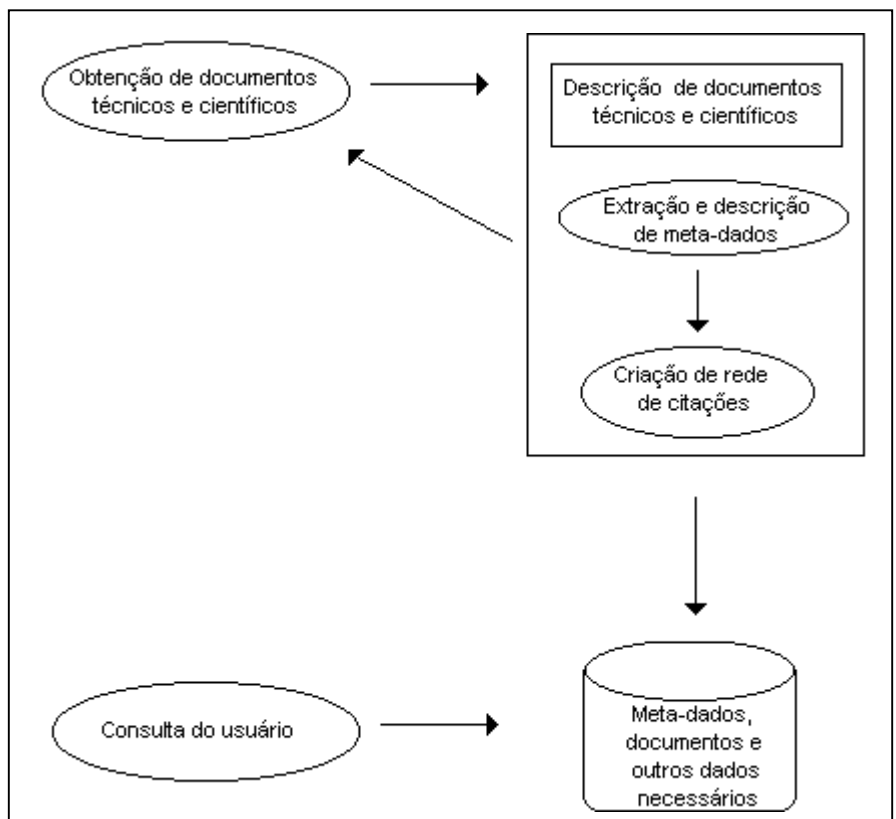
O problema a ser solucionado situa-se em um contexto mais geral que consiste na criação de uma biblioteca digital de documentos científicos na *Web*. O processo de alimentação do acervo desta biblioteca será o mais automático possível, podendo ser dividida em três fases:

- 1) Obtenção de documentos técnicos científicos por uma agente de rastreamento na Web;
- 2) Descrição dos documentos através de suas meta-informações, dividida em duas fases:
 - 2.1) Extração e descrição automática de metadados;
 - 2.2) Criação de uma rede automática de citações baseando-se nas descrições obtidas na fase anterior;
- 3) Recuperação de documentos em resposta às consultas de usuários.

As relações entre as fases do projeto podem ser vistas na Figura 1.

O presente documento tratará somente a fase 2.2, da criação de uma rede automática de citações, denominada Índice de Citação Autônomo (ICA). Um dos índices de citação mais populares disponível atualmente é o Science Citation Index [3]. O ICA tem a capacidade de melhorar a comunicação científica sobre vários aspectos:

- Revelando relacionamentos entre artigos;
- Ajudando a fazer correções nos trabalhos publicados;
- Identificando melhorias significativas nos trabalhos publicados;
- Ajudando a limitar a duplicação de pesquisas anteriores;
- Analisando linhas de pesquisa, e assim, identificando áreas emergentes da ciência;
- Encontrando onde e o quanto um artigo é citado.



Contudo, os índices de citações disponíveis e propostos na literatura científica atual dependem bastante da preparação humana ou de uma edição da informação. Por exemplo, Robert D. Cameron propôs um banco de dados universal de citações que faria a ligação entre todos os trabalhos já escritos [6]. Cameron descreveu um sistema no qual toda pesquisa publicada seria disponibilizada a qualquer estudante com acesso a Internet. O banco de dados incluiria ligação entre as citações e seria compreensível e atualizada. Esse sistema proposto por Cameron transferiria todo esforço manual associado à indexação das citações para os autores ou instituições que seriam solicitados para prover informação da citação em um formato específico.

Tais requisitos impostos aos autores são provavelmente um fator agravante ao impedimento da realização da proposta de Cameron. O Índice de Citação Autônomo (ICA), por outro lado, evita esses requisitos por automatizar completamente o processo de indexação de uma citação, sem requerer qualquer esforço extra de autores ou instituições. Adicionalmente, o ICA evidencia melhorias em outras tecnologias por extrair e fazer com que o contexto das citações tenha acesso mais facilitado.



4 Objetivo

O objetivo deste artigo é o desenvolvimento de uma ferramenta que crie uma rede de citações [16] automaticamente. O trabalho já realizado e relatado neste artigo, focalizou-se na melhor compreensão do problema, na especificação desta ferramenta e no desenvolvimento de um modelo de dados que constituirá a base para a concepção da rede de citações que permita identificar e recuperar os artigos citados e citantes de um determinado documento.

O ICA deverá permitir ao usuário uma forma de navegação em documentos teoricamente similares, relacionando os artigos com as palavras citadas. Algumas das principais tarefas disponibilizadas pelo Índice de Citação Autônomo, definidas com maiores detalhes na Seção 4.1, são: localizar artigos independentes da linguagem e das palavras no seu título ou em palavras-chaves e permitir navegação nos artigos mais antigos (a lista de artigos citados) e nos mais novos (quais artigos subsequentes citam o artigo corrente?) [17], já que o fato de um documento citar, ou ser citado por outro, define um tipo de relação interessante entre esses documentos.

5 Detalhamento do problema

Nesta seção será feita uma explicação mais detalhada do problema, esclarecendo quais serão os requisitos funcionais aos quais esse trabalho se dispõe a atender (subseção 4.1). Ela fará também uma explicação de como será feita a entrada dos dados para o sistema proposto (subseção 4.2), como esses dados serão manipulados (subseção 4.3) e, por fim, identificará os resultados gerados pelo trabalho, dando continuidade ao projeto no qual este trabalho está inserido (subseção 4.4).

Requisitos funcionais

Nessa subseção, serão apresentados alguns dos principais requisitos funcionais levantados e que o trabalho relatado nesse relatório se incumbirá de implementar.

Estatísticas

Estes requisitos funcionais são os mais importantes desse trabalho, ou seja, são as principais tarefas a serem providas, juntamente com o requisito definido na subseção 4.1.2.

O requisito sobre as Estatísticas pode ser desdobrado em dois, um fazendo referência aos artigos citados e outro referente aos artigos citantes. Exemplificando, em um artigo X, os artigos Y e Z são referenciados, nesse contexto, os artigos Y e Z são os chamados citados e o artigo X é denominado citante. Nas subseções 4.1.1.1 e 4.1.1.2 essas definições serão vistas com mais detalhes e o exemplo citado será utilizado para maior detalhamento do requisito.

Artigos citados

O requisito funcional denominado *Estatísticas sobre artigos citados* será aquele em que o usuário poderá saber quantas vezes o artigo Z foi citado como visto no exemplo acima, além de ter a informação de qual artigo o citou.



Artigos citantes

Esse requisito, *Estatísticas sobre artigos citantes*, se refere ao artigo denominado citante, no exemplo dado seria o artigo X. Através desse requisito, o usuário poderá saber qual artigo cita o artigo X e quantas vezes ele é citado.

Navegabilidade

O requisito *Navegabilidade* diz respeito à característica principal do ICA, juntamente com os requisitos 4.1.1.1 e 4.1.1.2. Pois através de um artigo, por exemplo, X, o usuário poderá além de saber quais artigos o cita e quais são citados por ele, navegar nessa rede de citações.

Tal navegação, poderá ser feita tanto para o passado no tempo, através da lista de artigos citados, quanto para o futuro, através da busca por artigos mais recente, entre os relacionados.

Contextualização

Outro requisito a ser disponibilizado e também de grande importância é o de *Contextualização*. Pois o usuário poderá visualizar o contexto com o qual um artigo foi citado. No exemplo dos artigos X, Y e Z, o usuário saberá que o artigo Z foi citado por X, e terá disponibilizado um pequeno trecho do texto do artigo no qual a referência a Z é feita.

Perfil do pesquisador

O projeto no qual esse trabalho está inserido terá um cadastro de pesquisadores (autores dos artigos indexados) e, nesse cadastro, o pesquisador poderá definir algumas áreas de pesquisa às quais ele se interessa. Com isso, o pesquisador poderá escolher, quando do seu cadastro, se ele quer receber alguma novidade do sistema.

A novidade dita acima diz respeito a novos artigos que forem sendo indexados pelo ICA. Com o sistema sabendo se um pesquisador tem interesse na área de bibliotecas digitais, por exemplo, quando um novo artigo for indexado pelo ICA e este contiver alguma referência ao assunto bibliotecas digitais, o pesquisador poderá receber uma notificação deste novo artigo ao entrar no sistema.

Aviso ao pesquisador

Através dessa funcionalidade, quando um novo artigo, por exemplo, X, for indexado pelo ICA e este artigo fizer referência a um outro artigo Y de um determinado autor, se esse autor for cadastrado no sistema, ele será notificado que seu artigo Y foi citado pelo artigo X.

Esse requisito funcional denominado *Aviso ao pesquisador* será uma boa maneira de um pesquisador saber que seu trabalho está sendo reconhecido.

Entrada de dados

Como já citado anteriormente, este trabalho faz parte de um projeto maior e, com isso, torna-se necessário o uso de uma interface entre a parte de extração e descrição automática de metadados e o ICA, fazendo com que os metadados gerados sirvam de alimentação para o modelo de dados detalhado na Seção 5.

Para a viabilização da interface supracitada, os metadados descritivos utilizados serão aqueles que fazem parte do padrão *DCMES* [5], sendo representados utilizando *RDF* (*Resource Description Framework*) [8] e expressados em *XML*



(*eXtensible Markup Language*) [9]. Tais metadados que servirão de entrada, serão obtidos através de arquivos.

Uma melhor explicação da definição de XML e de RDF será dada nas subseções 4.2.1 e 4.2.2, respectivamente.

XML – Extended Markup Language

XML é uma linguagem de marcação utilizada para a descrição de dados que é derivada da linguagem *SGML* (*Standard Generalized Markup Language*) [10], uma linguagem usada para a descrição de outras linguagens [11]. *XML* foi criada para auxiliar a troca de documentos eletrônicos na *Web*.

A descrição de dados em *XML* é restrita ao nível sintático e não há, a priori, qualquer significado nas marcações de *XML*. Uma aplicação que processe um documento *XML* deve conhecer o significado de cada *tag* do documento, já que não é definido qualquer conjunto de *tag*, a priori. Cada aplicação irá definir seu próprio conjunto de *tags* bem como a semântica destes.

RDF – Resource Description Framework

RDF é um “*framework*” para metadados cujo objetivo é promover a interoperabilidade entre aplicações que compartilham informações compreensíveis por máquinas na *Web* [12]. RDF não define qualquer semântica nem faz qualquer pressuposto sobre qualquer domínio de conhecimento, já que deve ser um mecanismo de descrição neutro que sirva para descrever recursos de qualquer campo do conhecimento [13].

O RDF é composto por três tipos de objetos: Recursos, Propriedades e a Tripla.

Um recurso é o que será descrito por uma expressão RDF. Todo recurso é identificado por um *URI* (*Uniform Resource Identifier*, incluindo aí o *Uniform Resource Locator - URL*).

Uma propriedade é qualquer característica utilizada para descrever um recurso. Em RDF um domínio de conhecimentos é definido via um *RDF Schema* [14]. É em um *RDF Schema*, portanto, que é definida a semântica e as características de uma propriedade. Uma aplicação que crie metadados em RDF e outra que utilize esses metadados devem utilizar o mesmo *Schema* para um funcionamento adequado.

Uma tripla é um recurso, uma propriedade e o valor da propriedade para um recurso. Uma tripla possui a forma <*sujeito, predicado, objeto*>. O significado de uma tripla pode ser resumido como o recurso (*sujeito*) possui a propriedade (*predicado*) com este valor (*objeto*). Um objeto pode ser tanto um outro recurso quanto um tipo primitivo definido por XML. Por exemplo, a tripla <“*http://www.dcc.ufmg.br/~igorleal*”, “*criador*”, “*Igor Campos Leal*” > teria o significado: Igor Campos Leal é o criador da página *http://www.dcc.ufmg.br/~igorleal*.

Dublin Core

O *Dublin Core Metadata Elements Set (DCMES)* [5] é uma ontologia criada pela *Dublin Core Metadata Initiative (DCMI)* [15] para a descrição de recursos *Web*. *DCMES* provê um vocabulário semântico para descrever as propriedades básicas de um recurso, por exemplo, o autor (*Creator*) e o título (*Title*).

O conjunto de metadados *Dublin Core* é formado por quinze elementos: *Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier,*



Source, Language, Relation, Coverage, Rights. Uma descrição de cada elemento pode ser encontrada em [5].

Todos os elementos relacionados acima são opcionais ao descrever um recurso. Além disso, não há restrição quanto ao número de vezes que cada elemento pode ser utilizado nem quanto ao tamanho do valor de cada elemento.

Manipulação dos metadados

Os metadados que serão gerados pela etapa precedente e que servirá de entrada para esse trabalho fazem parte do conjunto de metadados *Dublin Core*. Abaixo segue a lista dos metadados que serão utilizados pelo presente trabalho.

Elemento	Dados Extraídos
Title	Título do artigo.
Creator	Autor do artigo. Para cada autor identificado será gerado um metadado.
Subject	Conjunto de palavras chaves, separadas por ‘;’.
Description	Resumo do artigo (abstract).
Relation	Citações feitas no artigo.
Format	Formato do recurso, no caso postscript ou PDF.

A ferramenta que será implementada utilizará os dados acima citados para o povoamento do banco de dados a ser utilizado através do modelo de dados definido na Seção 5.

Resultados esperados

Como já dito na introdução desse documento, as etapas de um processo [21] podem ser definidas segundo a figura 2, mostrada abaixo.

Como, através da subseção 4.2, foi definida

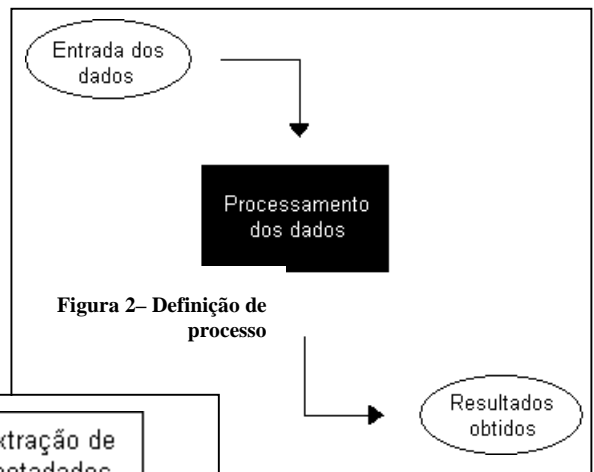
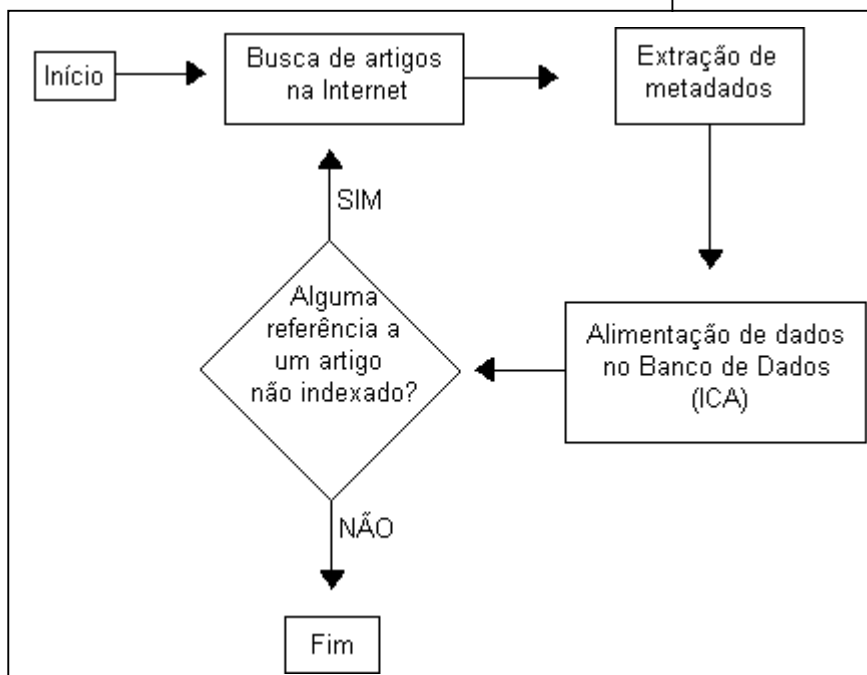


Figura 2- Definição de processo



como será feita a entrada dos dados, cabe ressaltar agora, quais serão os resultados obtidos, ou seja, a partir de algum processamento de dados, o que será gerado como saída. Essa saída, detalhada na

Figura 3 - Etapas



presente subseção, não se refere àquela com que o usuário do sistema se deparará, mas sim, àquela referente à alimentação de todo o projeto, já que a entrada para o atual trabalho provêm de um outro, sendo que o mesmo vale para a saída deste. Sendo assim, podemos definir um ciclo entre as etapas do projeto segundo o fluxograma da Figura 3.

Segundo este fluxograma, a saída será utilizada como entrada para a etapa de *Busca de artigos na Internet*, se for necessário. Essa interface será feita a partir dos dados (presentes ou faltantes) do Banco de Dados proposto na Seção 5. E, para que essa interface seja o mais eficiente possível, o ICA deve ser suficientemente eficaz na manipulação dos metadados, gerando uma carga de dados fidedigna ao texto extraído do artigo.

6 Modelagem de dados

Nesta seção será ilustrado como o sistema proverá os requisitos funcionais definidos na subseção 4.1, propondo um modelo de dados segundo a modelagem ER, definida na subseção 5.1. A partir do estudo feito, pôde-se escolher o modelo que melhor atende os requisitos levantados na subseção 4.1, e, com isso, a especificação, a análise e o desenho do modelo de dados escolhido serão apresentados nas subseções seqüentes.

Modelo ER

O Modelo Entidade-Relacionamento [20] é baseado em uma percepção do mundo como consistindo de uma coleção de objetos (entidades) e os relacionamentos entre esses objetos.

Uma entidade é um objeto que distingue de outros e que existe. Cada entidade é associada a um conjunto de atributos que a descreve, por exemplo, número de balanço para uma entidade *Conta*.

Um relacionamento é uma associação entre várias entidades, por exemplo, um relacionamento *Cliente_conta* associa um cliente com cada conta que o pertença.

Um conjunto de todas entidades e relacionamentos de mesmo tipo é chamado o conjunto entidade ou conjunto relacionamento. Outro elemento essencial para o diagrama ER é o mapeamento das cardinalidades, que expressa o número de entidades com a qual outra entidade pode ser associada via um conjunto relacionamento.

A estrutura lógica global de um banco de dados pode ser expressa graficamente por um diagrama ER:

- retângulos: representam conjuntos entidade;
- elipses: representam atributos;
- losangos: representam relacionamentos entre conjuntos entidade;
- linhas: fazem a ligação dos atributos aos conjuntos entidade e a ligação dos conjuntos entidade aos relacionamentos.

Comparação

A fim de encontrar uma modelagem de dados que possa mapear os requisitos levantados na subseção 4.1 para um diagrama ER, uma análise criteriosa de cada requisito deverá ser feita e será demonstrada em cada uma das subseções abaixo.



Requisito Estatísticas sobre artigos citados

Para prover este requisito, o sistema deverá ter armazenado no seu banco de dados o artigo em si, com todos os dados referentes a ele, bem como todas as suas referências bibliográficas, para que se possa levantar os artigos que são citados.

Para esse requisito, deverá ser criada uma entidade denominada *Artigo*, com todos os dados deste, como, por exemplo, título, data de publicação, resumo e texto completo. E para as referências bibliográficas deverá ser criada uma entidade *Referência*, com os atributos: sigla e detalhes da referência. Como exemplo, a referência “[20] *ELMASRI & NAVATHE, Fundamentals of Database Systems, 3rd ed., Benjamin/Cummings Pub., 1999.*” tem como sigla “20” e como detalhes “*Elmasri e Navathe, Fundamentals of Database Systems, 3rd ed., Benjamin/Cummings Pub., 1999.*”.

A entidade *Artigo* deverá se relacionar com a entidade *Referência*, relacionamento este denominado *Tem_as*, indicando todas as referências bibliográficas de um artigo. Além deste, a entidade *Artigo* deverá ter um outro relacionamento chamado *Faz_referência_aos*, mas para ela mesma. Pois a cada referência feita, o ICA deverá solicitar a busca das informações referentes a esse artigo referenciado e guardar as informações encontradas na entidade *Artigo*, como explicitado na Figura 3.

Requisito Estatísticas sobre artigos citantes

Para a viabilização dessa funcionalidade, as mesmas características do requisito anterior devem ser atendidas. Pois, sabendo-se quais artigos são citados, sabe-se quais são os citantes. E, com isso, as mesmas entidades com os mesmos atributos deverão ser criadas.

Requisito Navegabilidade

Para que esse requisito funcional seja atendido, a criação das entidades *Artigo* e *Referência* e os relacionamentos *Tem_as* e *Faz_referência_aos* são suficientes. Pois, como se saberá que o artigo X cita os artigos Y e Z, e, por conseqüência, os artigos que tanto Y quanto Z fazem referências, a navegação está disponibilizada.

Requisito Contextualização

Sobre o requisito *Contextualização*, pode-se perceber que ele é automaticamente possibilitado, pois sabendo-se qual artigo cita qual, tendo a sigla, o nome da citação feita e o texto completo do artigo, basta fazer uma pesquisa no texto do artigo em busca da sigla em questão. Com isso, deve-se extrair um pequeno trecho do texto no ponto em que a referência é feita.

Requisito Perfil do pesquisador

Para que seja disponibilizado o requisito funcional *Perfil do pesquisador*, o modelo deve representar os autores de cada artigo, e, para isso, uma entidade *Pesquisador* deverá ser criada. Essa entidade *Pesquisador* poderá conter algumas informações sobre cada autor de um determinado artigo, como por exemplo nome, e-mail e endereço na Internet.

Para saber qual é o perfil de um pesquisador, deve-se guardar a informação de suas áreas de interesse. Isso poderá ser descoberto através das palavras-chaves de seu artigo, pois se um autor escreve um artigo com determinadas palavras-chaves, isso significa que ele, o autor, tem interesse nessas áreas. Então, a criação de uma entidade denominada *Áreas_de_interesse* deve ser feita, bem como a criação de um

relacionamento entre as entidades *Pesquisador* e *Áreas_de_interesse*, denominado *Se_interessa_por*. De posse dessas informações, torna-se totalmente viável uma notificação ao pesquisador através de alguma mensagem do sistema quando este estiver em uso do pesquisador e quando algum novo artigo fizer referência a alguma de suas áreas de interesse.

Aproveitando a criação da entidade *Áreas_de_interesse*, será criado um relacionamento entre a mesma e a entidade *Artigo* para que seja possível armazenar no banco de dados todas palavras-chaves de um artigo. Tal relacionamento será chamado de *Contêm_as*. Dessa forma, as palavras-chaves serão utilizadas de duas formas diferentes: como a área de interesse do pesquisador e como palavra-chave do artigo. Além desse relacionamento, será criado um outro, chamado *Foi_escrito_por*, entre as entidades *Artigo* e *Pesquisador*, para que os autores de um artigo qualquer sejam armazenados.

Requisito Aviso ao pesquisador

Tendo armazenado em um atributo da entidade *Pesquisador* o e-mail de um pesquisador e fazendo uso do relacionamento *Foi_escrito_por*, entre as entidades *Artigo* e *Pesquisador*, torna-se possível a função de *Aviso ao pesquisador*. Já que se sabe quais são os autores de um determinado artigo e sabe-se quando este artigo é citado por algum outro artigo indexado pelo ICA, o autor em questão pode ser notificado por meio de um e-mail sobre essa citação a algum artigo de sua autoria.

ER Final

Abaixo segue o diagrama ER para a modelagem proposta na subseção anterior.

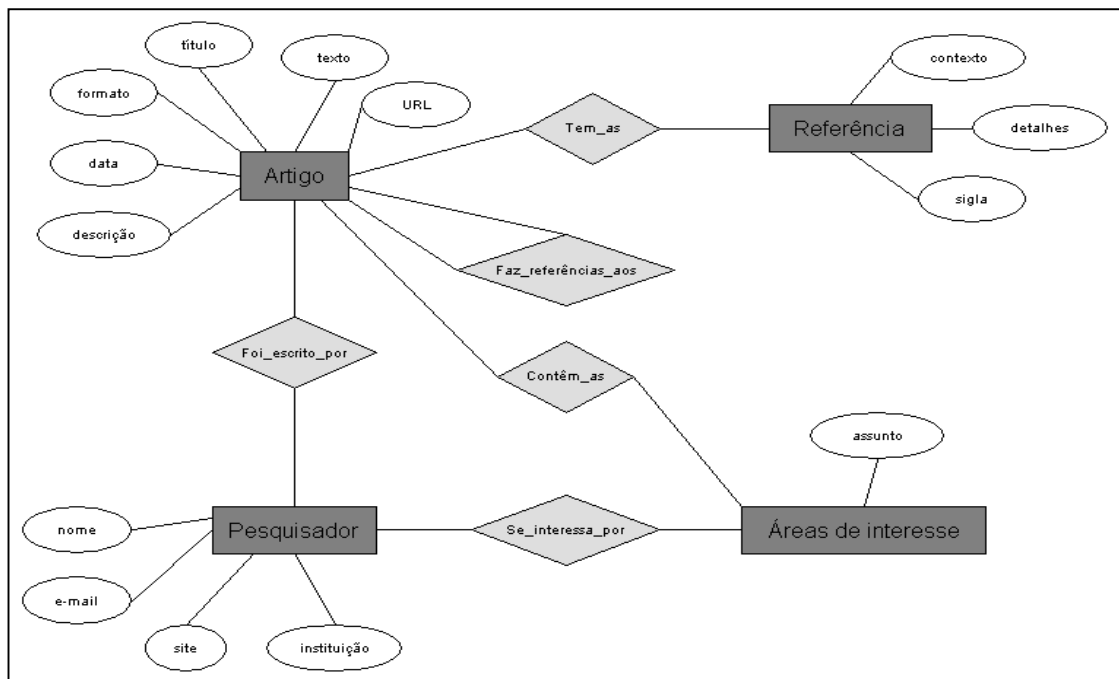


Figura 4 - Modelo ER



7 Análise de desempenho do ER

No sentido de verificar a carga de utilização do banco de dados modelado segundo a subseção 5.3, poderíamos estudar a possibilidade de aplicar Modelos Analíticos, por meio de cálculos matemáticos ou Modelos de Simulação, por meio de ações computacionais.

Contudo, essa tarefa deverá ser realizada em uma próxima fase do projeto, já que a carga para a simulação ainda não está disponibilizada e esta tarefa tomaria tempo excessivo, impedindo o andamento do trabalho.

8 Trabalhos relacionados

Algumas iniciativas nacionais e internacionais já vêm sendo feitas, como por exemplo o projeto SciELO, o programa Prossiga e o projeto CiteSeer. Nas subseções seqüentes, serão apresentados cada projeto e uma pequena explicação sobre suas atuações.

Projeto SciELO

O projeto SciELO [7], financiado pela Fapesp, é uma iniciativa nacional para promover o acesso livre a um documento eletrônico. Tal projeto tem hoje registrado cerca de 51 revistas científicas nacionais e latino-americanas, trata-se, portanto, de uma iniciativa centrada em produtos oriundos de editores científicos.

Programa Prossiga

Criado em 1995, o programa Prossiga [19] tem por objetivo promover a criação e o uso de serviços de informação na Internet voltados para as áreas prioritárias do Ministério da Ciência e Tecnologia, assim como estimular o uso de veículos eletrônicos de comunicação pelas comunidades dessas áreas.

Visando o alcance do objetivo proposto, o Prossiga, na área de ciência, tecnologia e informação, fortalece a presença da informação brasileira na Internet atribuindo-lhe maior visibilidade e acessibilidade e estimula a criação e consolidação de comunidades virtuais.

CiteSeer

CiteSeer (ou *Researchindex.org*) [18] é um sistema de indexação automática de citações de documentos científicos. O sistema faz uma análise detalhada das citações do documento com o objetivo de criar uma rede de citações de documentos científicos. É feita a recuperação do(s) autor(es) e título da citação, sendo possível verificar se duas citações, com apresentação diferente em dois documentos diferentes, referem a um mesmo documento.

9 Conclusão

Apesar de vivermos em um mundo dominado pelas tecnologias da informação e da comunicação, o Brasil não possui ainda uma forma sistematizada de registro e disseminação da sua produção científica. Falta, portanto, o desenvolvimento de sistemas computacionais que possibilitem tais benefícios ao pesquisador nacional.

Com o estudo mais detalhado do problema que este trabalho pretende solucionar, pôde-se concluir que, juntamente com os outros módulos, ele poderá atender a ação de registro e disseminação da produção científica nacional de forma



vantajosa em relação à solução tradicional de formação de uma rede de serviços de informação com essa finalidade, já que esta última possui custos proibitivos, pois são intensivos em mão-de-obra humana.

A formação de uma rede tradicional de serviços pressupõe o engajamento de instituições especializadas que coletariam, catalogariam e indexariam tais documentos. Estas atividades são realizadas por técnicos da área de Informação, mas, nos dias de hoje, não se pode mais pensar na implementação de tal solução face ao crescente volume de informação produzida. O momento tecnológico atual exige e permite a adoção de outras soluções que possam recuperar o tempo perdido e implantar no Brasil um mecanismo totalmente sistematizado. E o presente trabalho se insere no rol dessas soluções.

Através dessa inovação, o desenvolvimento a ser produzido traz ao país vários impactos, como por exemplo, a capacidade de tornar visíveis os resultados de pesquisas apoiadas e financiadas pelas agências de fomento; a criação de uma competência nacional no desenvolvimento de ferramentas para a construção de bibliotecas digitais; a criação de uma comunidade de autores com domínio em publicação eletrônica; a possibilidade da realização de estudos sobre a pesquisa científica e tecnológica no país; o conhecimento da produção da literatura científica nacional; e por fim, como consequência, a geração de indicadores que possam dar subsídios ao planejamento da ciência e da tecnologia no Brasil.

Neste contexto, pode-se perceber a grande importância do presente projeto, pois a biblioteca digital proposta terá funcionalidades não existentes nas bibliotecas tradicionais, estendendo o conceito de biblioteca muito além do conhecido hoje e, com isso, acelerando o desenvolvimento da ciência e da tecnologia e gerando mais riquezas para o país, no tocante do conhecimento nacional.

10 Referências Bibliográficas

- [1] LAWRENCE, S. & GILES, C. Searching the world-wide-web. *Science*, 280(5360):98, April 3, 1998.
- [2] LEAL, I. “Índice de Citação Autônomo”, Proposta de Trabalho de Projeto Orientado em Computação I.
- [3] GARFIELD, E., “Science Citation Index”, *Journal Citation Reports*, 1990, A bibliometric analysis of science journals in the ISI Data.
- [4] BAX, M.P., and KURAMOTO, H. “Pesquisa e Desenvolvimento de Ferramentas para a Captura, Tratamento e Recuperação de Literatura Científica na Web”, Projeto de pesquisa submetido ao CNPq em julho de 2001.
- [5] Dublin Core Data Set elements - <http://dublincore.org/documents/dces/>
- [6] CAMERON, R.D., "A Universal Citation Database As a Catalyst for Reform in Scholarly Communication," *First Monday*, Apr. 1997, http://www.firstmonday.dk/issues/issue2_4/cameron/index.html.
- [7] PACKER & LAERTE, A. & BIOJONE & ROCHA M. et al., “SciELO: uma metodologia para publicação eletrônica”, *Ci. Inf.*, 1998, vol. 27 n° 2. ISSN 0100-1965.
- [8] Semantic Web Activity: Resource Description Framework (RDF) - <http://www.w3.org/RDF/>
- [9] Extensible Markup Language (XML) - <http://www.w3.org/XML/>
- [10] Standard Generalized Markup Language - ISO 8879



- [11] BAX, M. P. - **Introdução às linguagens de marcas – Ciência da Informação, Brasília, v. 30, n. 1, p. 32-38, jan./abr. 2001**
- [12] **Frequently Asked Questions about RDF - <http://www.w3.org/RDF/FAQ>**
- [13] **Resource Description Framework (RDF) Model and Syntax Specification - <http://www.w3.org/TR/REC-rdf-syntax>**
- [14] **Resource Description Framework(RDF) Schema Specification - <http://www.w3.org/TR/1998/WD-rdf-schema/>**
- [15] **Dublin Core Metadata Initiative (DCMI) - <http://dublincore.org/>**
- [16] **GARFIELD, E., Citation Indexing: Its Theory and Application in Science, Technology, and Humanities, John Wiley & Sons, New York, 1979.**
- [17] **GILES, C. L. & BOLLACKER, K. D. & LAWRENCE, S. - “Digital libraries and autonomous citation indexing”, IEEE Computer, 32(6):67—71, 1999.**
- [18] **GILES, C. L. & BOLLACKER, K. D. & LAWRENCE, S. – Indexing and Retrieval of Scientific Literature - Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2–6, pp. 139–146, 1999.**
- [19] **Programa Prossiga - <http://www.prossiga.br>**
- [20] **ELMASRI & NAVATHE, Fundamentals of Database Systems, 3rd ed., Benjamin/Cummings Pub., 1999.**
- [21] **SCHISSATTI, M. L., “Uma metodologia de implantação de cartas de Shwart para o controle de processos”, cap. 2, dissertação de Mestrado da UFSC.**