MODELO DE MINERAÇÃO DE DADOS EM BASES DE DADOS ACADÊMICAS

RENAN MONTEIRO DA SILVA

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

MODELO DE MINERAÇÃO DE DADOS EM BASES DE DADOS ACADÊMICAS

RENAN MONTEIRO DA SILVA

ORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGEE.DM-625/16
BRASÍLIA/DF, 2016

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

MODELO DE MINERAÇÃO DE DADOS EM BASES DE DADOS ACADÊMICAS

RENAN MONTEIRO DA SILVA

DISSERTAÇÃO DE MESTRADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE.

APROVADA POR:

RAFAEL TIMÓTEO DE SOUSA JÚNIOR, Dr., ENE/UNB
(ORIENTADOR)

FLÁVIO ELIAS GOMES DE DEUS, Dr., ENE/UNB
(EXAMINADOR INTERNO)

SÉRGIO FARIAS DE ALBUQUERQUE, Dr., PNSOFT
(EXAMINADOR EXTERNO)

Brasília, 12 de abril de 2016.

FICHA CATALOGRÁFICA

SILVA, R. M.
Modelo de mineração de dados em bases de dados acadêmicas [Distrito Federal] 2016.
xvii, 111p, 210 x 297 mm (ENC/FT/UnB, Mestre, Engenharia Elétrica, 2016).
Dissertação de Mestrado – Universidade de Brasília. Faculdade de Tecnologia.
Departamento de Engenharia Elétrica.

1. Sistemas Distribuídos	2. Mineração de Dados
3.Clusterização	4.Aprendizado de Máquina
I. ENC/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

SILVA, R. M.. (2016). Modelo de mineração de dados em bases de dados acadêmicas. Dissertação de Mestrado em Engenharia Elétrica, Publicação PPGEE.DM-625/16, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 111p.

CESSÃO DE DIREITOS

AUTOR: Renan Monteiro da Silva.

TÍTULO: Modelo de mineração de dados em bases de dados acadêmicas. .

GRAU: Mestre ANO: 2016

Renan Monteiro da Silva
Rua Babaçu Lote 3 Condomínio Art Life Design Apartamento 1304 - DF.
Brasília-DF – Brasil – 71.928-000
Celular: (61) 9197-1523 / e-mail: renanmonteirosilva@gmail.com

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado forças e motivação para alcançar meus objetivos e por ter me amparado durante toda essa jornada.

Agradeço a minha esposa pela compreensão e pelo companheirismo durante todas as etapas do meu mestrado.

Agradeço ao meu orientador Professor Rafael T. de Sousa Jr., por suas orientações, ensinamentos e oportunidades oferecidas durante esse processo.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pelo auxílio através da bolsa provida para os fins desse trabalho de mestrado.

Agradeço ao Ministério do Planejamento, Orçamento e Gestão – MPOG, especificamente à Secretaria de Relações do Trabalho – SRT, pela oportunidade de atuar em um projeto relacionado à linha de pesquisa nesse trabalho.

Agradeço também aos colegas do Laboratório de Tecnologia da Tomada de Decisão – LATITUDE, em especial a Fábio Buiati, Fábio Lúcio Lopes de Mendonça e Valério Aymoré Martins, pelo apoio fornecido durante o desenvolvimento dessa pesquisa.

Por fim agradeço a minha família pelo apoio e suporte, em especial a meu irmão Henrique Monteiro Alves pela ajuda fornecida na confecção das ilustrações desse trabalho.

RESUMO

Título: **Modelo de mineração de dados em bases de dados acadêmicas**
Autor: **Renan Monteiro da Silva**
Orientador: **Rafael Timóteo de Sousa Júnior**

No campo das comunidades de pesquisa existe uma série de bases de dados que proveem informações interessantes sobre publicações resultantes da pesquisa, incluindo títulos de artigos, autores, palavras-chave, citações, índices, veículos de publicação (revistas, livros, conferências e os tipos de eventos mais importantes) e assim por diante. Exemplos de tais bases de dados são Google Scholar, CiteSeerX, DBLP, Microsoft Academic, Thomson Reuters Web of Science, entre outros.

No entanto, essas bases de dados globais ainda carecem de serviços que possam ser usados na procura por comunidades ou agrupamentos. Uma comunidade pode ser definida como um grupo de entidades, nesse caso autores e/ou universidades, que compartilham atributos ou relacionamentos semelhantes.

Neste trabalho é proposto um modelo de mineração e análise das informações contidas nessas bases de dados acadêmicas. A análise dessas informações apresentadas nos resultados visa à descoberta das universidades, autores e assuntos mais significativos dentro do contexto dos dados minerados.

Para isso foi feito um estudo de caso utilizando as informações contidas nas bases de dados do CiteSeerX e do DBLP como ponto de partida para a criação de um modelo genérico com o objetivo de ser aplicável a qualquer base de dados acadêmica.

No estudo de caso é feita uma extensa mineração nas bases de dados do CiteSeerX e do DBLP, a partir dessa etapa é feita a migração e tratamento dos dados originais obtidos para o modelo genérico proposto neste trabalho.

Com o modelo preenchido são aplicados os algoritmos e instruções para geração dos resultados que são subdivididos em três diferentes categorias: clusters, rankings e comunidades de relacionamento.

A partir dos resultados são investigadas as tendências atuais na colaboração entre autores e institutos educacionais usando as bases de dados do CiteSeerX e do DBLP. Com a obtenção das informações disponíveis foram construídos várias comunidades e agrupamentos usando as técnicas de clusterização existentes.

ABSTRACT

Title: **Data mining model in academics databases**
Author: **Renan Monteiro da Silva**
Supervisor: **Rafael Timóteo de Sousa Júnior**

In the field of the research community, several databases such as Google Scholar, CiteSeerX, DBP, Microsoft Academic, Thomson Reuter´s Web of Science among others provide interesting information about authors, citations, indexes, most relevant venues types and so on.

However, those global databases have limitations, especially in finding communities or clusters. A community can be defined as a group of entities, in this case authors and/or universities that share similar properties or relations.

In this work, it is proposed a model of data mining and analysis of the obtained information in these academics databases. The analysis of the presented information in the results aims the discovery of the universities, authors and subjects most significant inside the context of the mined data.

Thus a study case was realized using the CiteSeerX database as the start point for creating a generic model in order to be applied in any academic database.

In the study case an extensive data mining was performed in the CiteSeerX database, as well as the migration and treatment of the original data obtained for the generic model proposed in this work.

With the model data filled the proposed algorithms and the code instructions were applied for the generation of the results which are subdivided in three different categories: clusters, rankings and relationship communities.

From the results, the work is validated by showing the current trends in the collaboration between authors and educational institutes, using the CiteSeerX dataset. By mining the available information, several communities and clusters are revealed using the proposed techniques.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE ACRÔNIMOS

API	*Application Programming Interface*
ARFF	*Attribute-Relation File Format*
DBLP	*Database Systems and Logic Programming*
BD	*Banco de Dados*
DM	*Data Mining*
HTML	*Hypertext Markup Language*
HTTP	*Hypertext Transfer Protocol*
IA	*Inteligência Artificial*
KSCD	*Korea Science Citation Database*
OAI	*Open Archive Initiative*
PDF	*Portable Document Format*
RDF	*Resource Description Framework*
SGBD	*Sistema de Gerenciamento de Banco de Dados*
SCI	*Science Citation Index*
SQL	*Structured Query Language*
XML	*eXtensible Markup Language*
WEKA	*Waikato Environment for Knowledge Analysis*

ARTIGOS VINCULADOS A ESTE TRABALHO

1. SILVA, R. M., BUIATI, F & DE SOUSA, R. T. Author Clustering Analysis on CiteSeerX Research Data. The Seventh International Conference on Information. 2015.
2. SILVA, R. M., BUIATI, F & DE SOUSA, R. T. Author Clustering Analysis Mixing Information from Structured Publication Databases and Unstructured Web Data. INFORMATION-An International Interdisciplinary Journal. 2015.
3. MARTINS, V. A. ; SILVA, R. M. ; SOUSA JR, R. T. ; HOLANDA, M. T. . Improving the codeplex PHP OData producer with an automatic extensible workaround model and data aggregation extension support. In: 11th International Conference Applied Computing 2014, 2014, Porto. Proceedings of.... Lisboa: IADIS, 2014. v. 1. p. 37-44.

1. INTRODUÇÃO

A mineração de informações de grandes bases de dados científicas tem sido usada para várias aplicações em diversas áreas de avaliação e gestão, além dos domínios próprios da ciência, ensino e pesquisa. Essas bases de dados permitem que informações importantes sejam recuperadas, especialmente aquelas referentes a citações, instituições acadêmicas, editores, autores.

Pesquisadores acadêmicos, por exemplo, dependem da análise das citações para usá-las como ferramenta para avaliar o impacto das publicações em cada área de pesquisa, e assim obter um insumo valioso para a criação de suas pesquisas.

As conferências e revistas são extremamente importantes nesse sentido, pois acumulam a grande maioria de publicações, e abrangem um grande número de estudos. Outra forma interessante de pesquisa é a análise sobre as comunidades formadas entre autores e universidades.

A descoberta dessas comunidades e a construção dessa base de informações pode ser aplicada em várias situações. Por exemplo, um estudante que está buscando uma instituição para a realização do seu doutorado pode encontrar nessas bases científicas informações relevantes sobre áreas com maior número de publicações, relações com outras instituições, tópicos de interesse, entre outros. Nesse contexto, um aluno que está buscando uma universidade para fazer o doutorado na área de Mineração de Dados pode encontrar um ranking das universidades e departamentos que mais publicam nessa área, assim como o nome dos pesquisadores chefes. Além disso, outro fator importante é a relação entre essas universidades, os grafos de interesses, etc.

Dada a importância e utilidade dessas informações, nesta dissertação os dados oferecidos pela biblioteca CiteSeerX e do DBLP foram usados para a construção de comunidades de interesses por autores e por áreas. Além dos dados disponibilizados pela biblioteca também foi necessário buscar informações complementares para auxiliar no processo de construção da base de informações resultante.

Feita a coleta dos dados, foram necessárias várias outras etapas para tratamento e pré-processamento dos dados para a construção dos clusters, dos rankings e do restante dos resultados gerados por esse trabalho.

1.1. MOTIVAÇÃO

Apesar de existirem vários trabalhos que exploram as informações disponibilizadas pelas bases de dados acadêmicas, eles não tratam das redes de comunidades formadas por instituições acadêmicas e autores no trabalho de criação e publicação de material científico.

Tampouco esses trabalhos exploram diferentes bases de dados acadêmicas com o objetivo de criar modelos genéricos que visem atender diversas entradas de dados.

Outro fator importante é que esta dissertação usa como insumo bases de dados acadêmicas que têm como foco trabalhos científicos relacionados às áreas de engenharia e tecnologia da informação.

Para aprimorar os resultados, este trabalho, além de fazer a extração, transformação e carga dos dados das bases de dados acadêmicas utilizadas também faz uma mineração de dados que visa enriquecer as informações obtidas agregando valor aos modelos e resultados gerados.

No entanto, os modelos construídos são genéricos e podem ser aplicados a qualquer área de atuação, possibilitando assim uma análise mais ampla dos resultados gerados a partir das entradas de dados suportadas.

1.2. OBJETIVOS DO TRABALHO

O presente trabalho almeja criar modelos e insumos que venham a auxiliar no processo de construção de novos materiais científicos. Nesse contexto os objetivos esperados por essa pesquisa são:

1.2.1 OBJETIVO GERAL

O objetivo geral deste trabalho é a criação de um modelo abrangente que contemple todo o processo de obtenção, tratamento e geração de resultados a partir de informações provindas de diversas bases de dados acadêmicas para possibilitar a visualização agregada das informações e a geração de conhecimento a partir delas.

1.2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho são as geração dos seguintes resultados a partir a aplicação do modelo criado:

- Clusters com a colaboração entre os autores;
- Clusters com a colaboração entre universidades;
- Rankings com os tópicos mais publicados por instituição;
- Identificação das instituições acadêmicas e os autores com maior número de publicações dada uma entrada de dados;
- Identificação os assuntos mais publicados em revistas;
- Identificação os assuntos mais publicados em conferências.

1.3. METODOLOGIA

A proposta de pesquisa desta dissertação foi dividida em três fases com o intuito de promover o melhor entendimento da pesquisa desenvolvida. Com o uso dessa metodologia este estudo visa ao aprofundamento do tema proposto e dos problemas em aberto relacionados:

Fase 1: Realizar a revisão bibliográfica para identificar os trabalhos relacionados ao tema, identificar os assuntos ainda não tratados e os problemas em aberto.

Fase 2: Implementar um estudo de caso para tratar os assuntos e os problemas apresentados na Fase 1;

Fase 3: Apresentar os resultados obtidos após a implementação do estudo implementado na Fase 2, realizar análises nos resultados e identificar as contribuições geradas pela pesquisa.

1.4. CONTRIBUIÇÕES DO TRABALHO

Este trabalho busca trazer as seguintes contribuições:

- Criação de um modelo que possa ser usado a partir de diferentes entradas de dados provindas de diferentes bases de dados bibliométricas.

- Possibilitar a partir da aplicação do modelo:
    - A identificação dos assuntos mais abordados em publicações de conferências e revistas;
    - A identificação dos principais autores dos assuntos mais abordados;
    - O mapeamento de uma rede de colaboração entre instituições que trabalhem em conjunto;
    - O mapeamento de uma rede de co laboração entre autores que trabalhem em conjunto;
    - A identificação das instituições com maior relevância em número de publicações nos assuntos mais abordados.

1.5. ORGANIZAÇÃO DO TRABALHO

Este trabalho foi dividido em cinco capítulos, começando por esta introdução para contextualizar o problema a facilitar o entendimento da pesquisa. A organização dos demais capítulos é descrita a seguir:

O Capítulo 2 apresenta os conceitos das técnicas aplicadas para criação dos clusters e rankings. E também apresenta os trabalhos correlatos e suas diferentes abordagens e conclusões para o problema em comum.

Capítulo 3 apresenta a proposta do modelo para clusterização, assim como as etapas necessárias para criação dos clusters, desde a captura dos dados, tratamento dos dados, pré-processamento até a etapa de geração dos clusters e rankings.

O Capítulo 4 apresenta os resultados alcançados com a aplicação dos modelos propostos no capítulo anterior.

O Capítulo 5 apresenta a conclusão deste trabalho. Nele, os resultados encontrados são sintetizados. E por fim são assinalados os caminhos futuros para a sequência deste trabalho.

2. REVISÃO BIBLIOGRÁFICA E ESTADO DA ARTE

Este capítulo trata dos principais conceitos de extração, transformação e carga, mineração de dados, clusterização e de suas técnicas relacionadas para o mapeamento de grupos semelhantes usando grandes massas de dados.

Contudo o foco deste trabalho se aplica ao contexto de publicações acadêmicas relacionadas às áreas de engenharia elétrica e tecnologia da informação, com o objetivo de dar suporte ao entendimento dos assuntos e problemas abordados neste trabalho.

2.1. EXTRAÇÃO TRANSFORMAÇÃO E CARGA

Segundo Liu (2009) o processo de extração, transformação e carga (ETL) é responsável pelas operações que acontecem em segundo plano em uma arquitetura de data warehouse. Em uma descrição de alto nível o processo de ETL primeiramente extrai o dado de uma determinada fonte de informação e posteriormente o armazena em outra fonte de dados, que pode ser um sistema legado, arquivos de qualquer formato, *stream* de dados, qualquer tipo de documento, entre outros.

Ainda segundo Liu (2009) em um processo de ETL apenas os dados diferentes dos que foram extraídos nos processos anteriores devem ser inseridos, atualizados ou deletados em uma execução posterior do mesmo processo.

Liang (2007) afirma que o processo de ETL referencia três funções separadas que quando usadas em conjunto formam um único procedimento computacional. Primeiramente a função de extração efetua a leitura de uma fonte de dados específica e extraí os dados desejados. Posteriormente a função de transformação trabalha com os dados adquiridos, efetuando os devidos tratamentos e transformações, e os converte para o formato final desejado. Por fim a função de carga cria o resultado esperado de acordo com o formato requerido, que pode ser em forma de base de dados, arquivos, documento, entre outros.

2.2. MINERAÇÃO DE DADOS

Witten (2001) afirma que a mineração de dados pode ser definida como o processo de descoberta de padrões em uma massa de dados. Esse processo pode ser automático ou semiautomático. Os padrões descobertos devem ter algum significado no sentido de conduzir a alguma conclusão significativa.

Ainda segundo Witten (2001) a utilização de padrões permite fazer a predição complexa sobre os novos dados. Em relação aos padrões existem dois extremos, o conceito de caixa preta (*black box*) cujo o interior é incompreensível e o conceito de caixa transparente cuja a estrutura de construção dos dados revela o padrão. Ambos os padrões são capazes de fazer boas predições. A diferença é se os padrões que são minerados podem ser representados na forma de uma estrutura que possa ser examinada, fundamentada e usada para auxiliar na tomada decisões futuras.

2.3. TÉCNICAS DE CLUSTERIZAÇÃO

Para a implementação de um cluster em uma massa de dados é necessário executar pelo menos três etapas que consistem na definição do algoritmo a ser usado, da distância de medida e do tipo de cluster desejado. A Figura 2.1 exemplifica o diagrama com as combinações que podem ser aplicadas de acordo com a entrada de dados a ser usada para gerar o cluster e com o tipo de cluster desejado.

Nas seções a seguir são explicadas mais profundamente as três características para a implementação de um cluster.

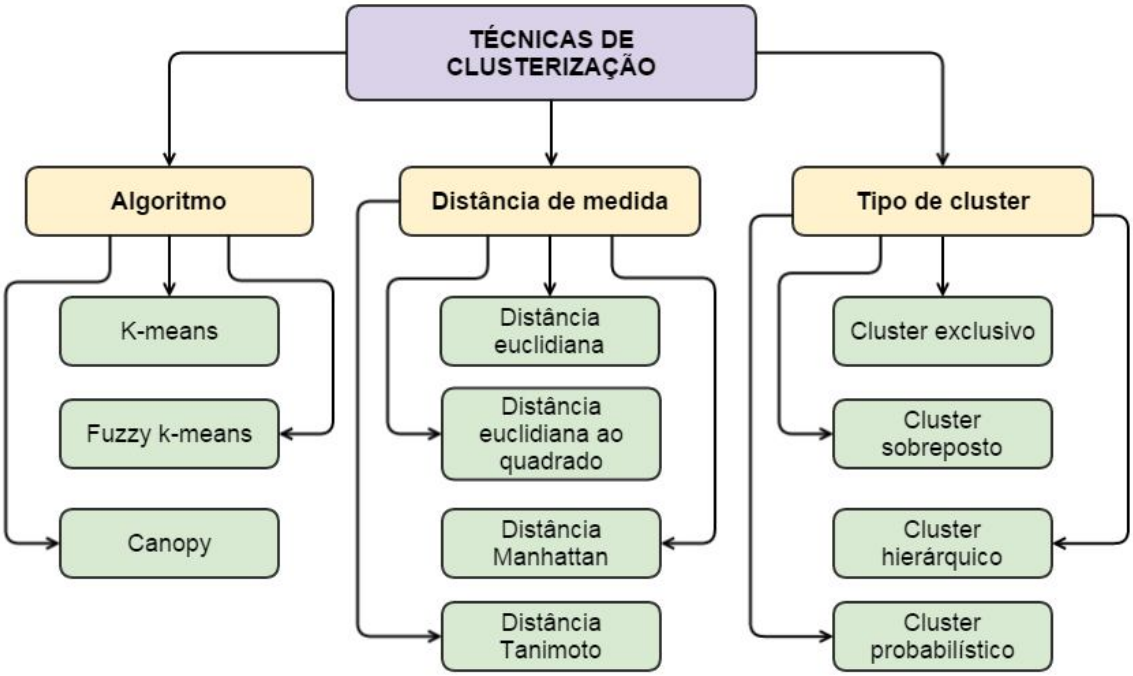

Figura 2.1 - Diagrama das técnicas de clusterização

2.3.1 TIPOS DE CLUSTER

A seguir são apresentados os principais tipos de clusters. Cada um tem sua aplicabilidade aprimorada para determinado contexto.

2.3.1.1. Cluster exclusivo

Segundo Jain et al (1999) o cluster exclusivo aloca cada padrão para um agrupamento distinto durante sua execução e como sua saída de dados.

Esse tipo de cluster, também conhecido como cluster rígido (hard cluster) não permite que um item pertença a mais de um cluster ao mesmo tempo, pertencendo a apenas um item, exclusivamente.

A Figura 2.2 apresenta a visualização do cluster exclusivo.

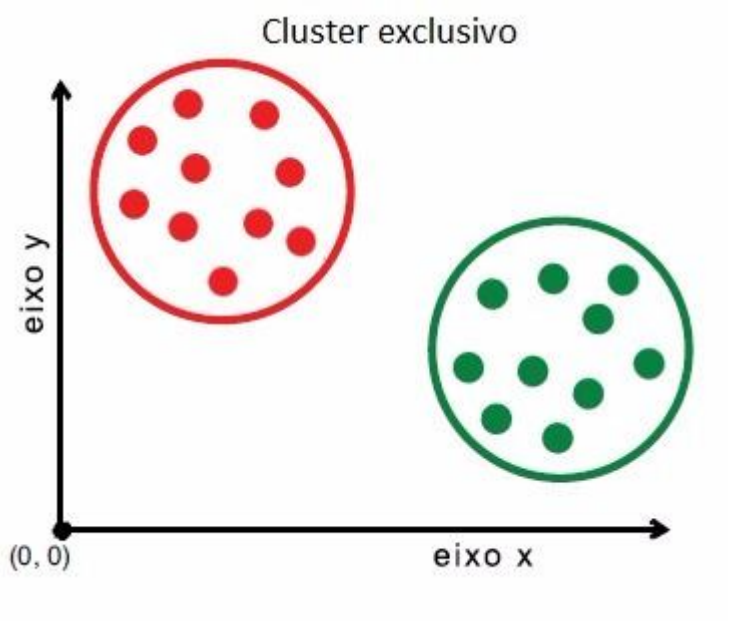

Figura 2.2 - Cluster exclusivo

2.3.1.2 - Cluster sobreposto

Essa abordagem é contrária à anterior por permitir que um ou mais itens pertençam a um ou mais clusters ao mesmo tempo, gerando assim subconjuntos como pode ser visto na Figura 2.3, que apresenta um exemplo básico de um cluster sobreposto.

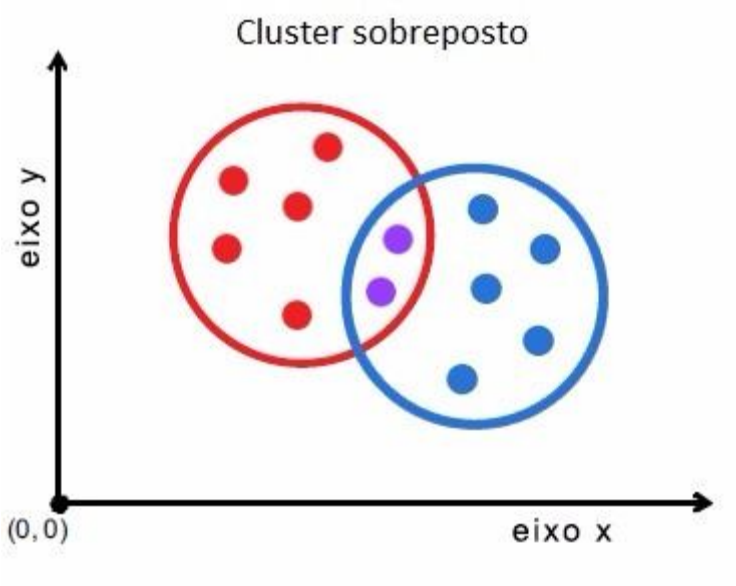

Figura 2.3 - Cluster sobreposto

Jain et al (1999) afirma que o método do cluster sobreposto atribui graus de participação em vários clusters para cada padrão de entrada. E que um cluster sobreposto pode ser convertido em um cluster exclusivo atribuindo cada padrão ao cluster com a maior concentração de valores.

2.3.1.3 - Cluster hierárquico

Essa técnica permite a criação de um agrupamento de itens em uma estrutura hierárquica, se assemelhando a uma árvore binária. É amplamente usada quando o cluster é muito grande para ser analisado. Nesse caso usando um cluster hierárquico é possível reduzir o seu tamanho agrupando os clusters menores em grupos.

Ainda segundo Jain et al (1999) o cluster hierárquico gera um dendograma representando o agrupamento de níveis de padrões e similaridade a cada mudança de agrupamento.

Figura 2.4 apresenta a visualização de um cluster hierárquico.

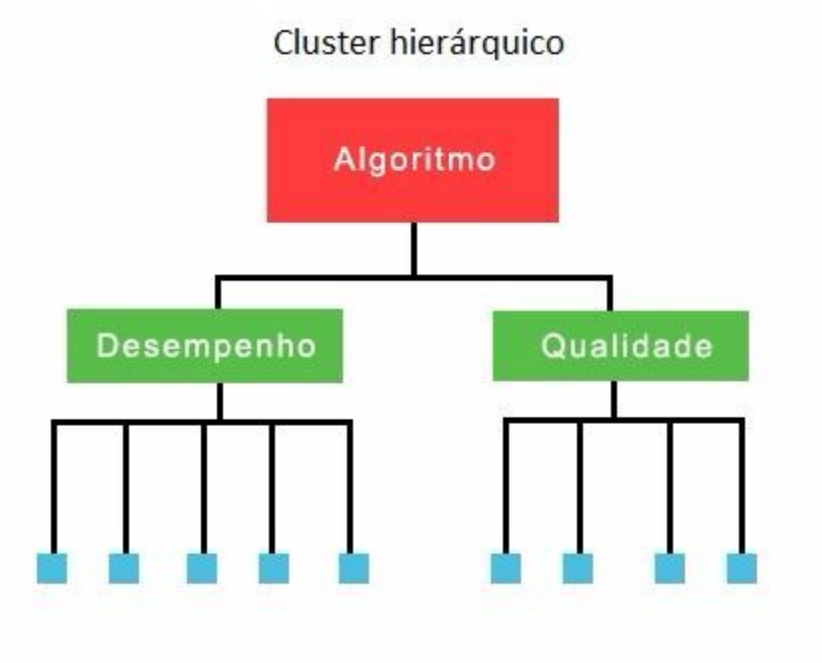

Figura 2.4 - Cluster hierárquico

2.3.1.4 - Cluster probabilístico

Owen et al (2012) afirma que esse tipo de cluster tenta aplicar um modelo probabilístico sobre a entrada de dados ajustando os parâmetros do modelo para agrupar os dados corretamente. Como um ajuste correto raramente acontece, então os algoritmos dão uma porcentagem correspondente ou um valor probabilístico que indicam o quão bem o modelo se ajustou ao cluster.

A Figura 2.5 ajuda a entender o funcionamento desse tipo de cluster. Nela são usadas uma forma elíptica e outra mais arredondada para serem aplicadas aos clusters, e descobrir os eixos e o centro dos clusters.

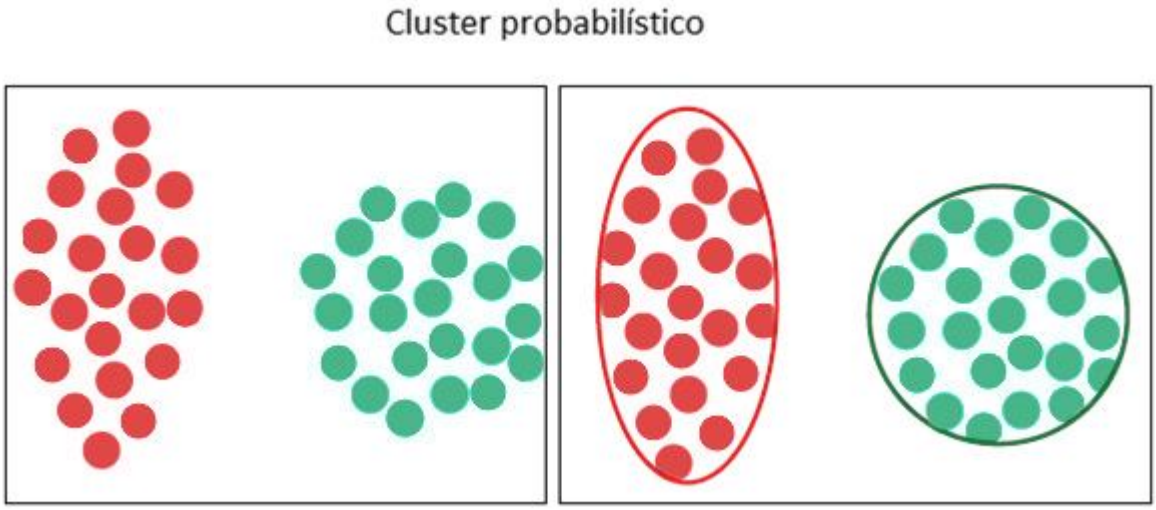

Figura 2.5 - Cluster probabilístico

2.3.2 - ALGORITMOS

Jain e Dubes (1988) declaram que o algoritmo de cluster é usado para agrupar grupos de objetos, ou grupos de dados, baseado nos índices de proximidade entre os pares de objetos.

O algoritmo aplicado para geração de um cluster é um fator essencial para alcançar resultados de qualidade. Sua escolha deve considerar o tipo de entrada de dados e o tipo de cluster que se deseja obter. Cada algoritmo é otimizado para granjear melhor desempenho e qualidade para determinada fonte de dados. Existem algoritmos que são mais apropriados para fontes de dados do tipo texto e outros que trabalham melhor com entradas numéricas.

O tipo de cluster requerido também influencia na escolha do algoritmo. Por exemplo, para se gerar um cluster sobreposto seria indicado o uso do algoritmo Fuzzy K-means, que é implementado de forma a gerar cluster desse tipo.

A seguir são detalhados alguns dos principais algoritmos de cluster e sugeridos os melhores contextos em que eles poderiam ser aplicados considerando a entrada de dados, tipo de cluster desejado e a qualidade do cluster:

2.3.2.1 - K-means

O algoritmo K-means foi criado há mais de cinquenta anos por Stuart Lloyd. Inicialmente projetado como uma técnica de modulação por código de pulsos, mais tarde adaptada para ser usada para o agrupamento de grupos. Tem uma implementação simples, no entanto é amplamente utilizado em várias áreas científicas.

O funcionamento do algoritmo acontece em dois passos O primeiro passo encontra os pontos mais próximos de cada centroide e os atribuem para um cluster específico. No segundo passo os centroides são recalculados usando a média das coordenadas de todos os pontos naquele cluster.

Como entrada o algoritmo recebe o número de clusters (K), a distância de medida a ser utilizada para calcular os pontos dentro do cluster e a fonte de dados. A Figura 2.6 apresenta o pseudocódigo do algoritmo.

```
//Lê os objetos
N = Número de objetos
K = Número de clusters

//Seleciona os primeiros objetos K como os centros iniciais dos clusters
objetos[N]: matriz de objetos
clusters[K]: matriz dos clusters
membros[N]: matriz dos membros

//Para cada objeto encontra o cluster mais próximo
for i = 0 ate N-1
    for j = 0 ate k-1
        distancia = objetos[i] - clusters[j]
        se distancia < distancia_minima
            distancia_minima = distancia
            n = j
    se membros[i] <> n
        membros[i] = n
    novo_cluster[n] = novo_cluster[n] + objetos[i]
    tamanho_novo_cluster[n] = tamanho_novo_cluster[n] + 1
//Recalcula os centroids
for j = 0 ate K-1
    clusters[j][*] = novo_cluster[j][*]/tamanho_novo_cluster[j]
    novo_cluster[j][*] = 0
    tamanho_novo_cluster[j] = 0
```

Figura 2.6 - Pseudocódigo do algoritmo k-means.

Manning (2008) afirma que idealmente em um cluster que usa o algoritmo *k-means* a esfera com o centroide fica no centro de gravidade entre os pontos do cluster, ou seja, os pontos ficam perfeitamente alinhados em torno do centroide. Ele também afirma que esse tipo de cluster não deveria se sobrepor, o que o caracterizaria como um cluster exclusivo.

2.3.2.2 - Fuzzy K-means

Segundo Owen et al (2012) o algoritmo *fuzzy k-means* é uma forma difusa da implementação do *k-means*, ao invés de produzir um cluster exclusivo essa implementação tenta gerar um cluster sobreposto. Esse algoritmo pode ser visto como uma extensão do *k-means*.

Também conhecido como *fuzzy c-means* na comunidade acadêmica, essa implementação trabalha de forma semelhante ao *k-means*, se diferenciando no conceito de

que um ponto pode estar associado a um ou mais clusters gerando assim um cluster sobreposto.

2.3.2.2 - Canopy

A geração *canopy*, também conhecida como *canopy clustering*, é uma técnica rápida de aproximação de agrupamento. É usada para dividir a entrada de dados em pontos de um cluster sobreposto. O algoritmo *canopy* tenta estimar o centroide do cluster mais próximo usando dois limiares de distância. (Owen et al 2012).

Esse algoritmo fornece o número de clusters gerados sem a necessidade de fornecê-los como parâmetro, como acontece com o *k-means*. Também tem como característica a capacidade de criar clusters de forma muito rápida, especialmente por fazer apenas uma interação sobre a entrada de dados, e é indicado para grandes massas de dados.

A Figura 2.7 a seguir apresenta o pseudocódigo do algoritmo *Canopy*.

```
//Lê os objetos
N = Número de objetos
objetos[N]: matriz de objetos
canopy[]: matriz dos clusters
centers[]: centros dos clusters

//Seleciona randomicamente um ponto da matriz de objetos para o centro do canopy
for i = 0 ate N-1
    if(canopy = null)
        centers[i] = objetos[i]
        objetos.removeIndex[i]
    else
        //Percorre a lista de objetos formando os clusters
        for c = 0 ate canopy-1
            if(similaridade(i,c)>T1)
                canopy[i] = objetos[i]
                objetos.removeIndex[i]
            if(similaridade(i,c)>T2)
                canopy[i] = objetos[i]
            if(similaridade(i,c)<T2)
                centers[i] = objetos[i]
                objetos.removeIndex[i]
            if(objetos = null)r
                return
```

Figura 2.7 - Pseudocódigo do algoritmo Canopy

Pelo fato de fazer apenas uma interação sobre a entrada de dados o desempenho é aprimorado, porém o custo dessa velocidade é a possibilidade dos clusters gerados não apresentarem uma precisão adequada.

2.3.3- MEDIDAS DE DISTÂNCIA

Segundo Qian (2004) as distâncias de medida são uma parte importante da modelagem dos vetores. Entre todas as distâncias de medidas que são propostas na literatura, algumas têm comportamentos bastante semelhantes nos cálculos de similaridade, enquanto outras se comportam de forma diferente. Entender a relação entre as distâncias de medida pode ajudar a escolher a distância de medida mais apropriada para determinada aplicação.

2.3.3.1 - Distância euclidiana

A distância euclidiana é calculada pela raiz quadrada da soma da diferença dos atributos ao quadrado, como pode ser visto na Equação 2.1.

$$dist(E_i, E_j) = \sqrt{\sum_{l=1}^{M}(x_{il} - x_{jl})^2}$$

Equação 2.1 - Distância euclidiana

Segundo Witten et al (2011) diferentes atributos são medidos em diferentes escalas, logo se a fórmula da distância euclidiana for aplicada diretamente pode gerar um completo ofuscamento sobre os atributos que se encontram em menor escala. Consequentemente é comum que todos os atributos sejam normalizados com valores entre 0 e 1 para o cálculo da distância.

2.3.3.2 - Distância euclidiana ao quadrado

A distância euclidiana ao quadrado é o quadrado da distância euclidiana, ela é calculada pelo quadrado da soma da diferença dos atributos, como apresentado na Equação 2.2.

$$dist(E_i, E_j) = \sum_{l=1}^{M} (x_{il} - x_{jl})^2$$

Equação 2.2 - Distância euclidiana ao quadrado

2.3.3.3 - Manhattan

A distância *Manhattan*, também conhecida como *City block*, é calculada diferentemente da distância euclidiana, que forma uma linha reta entre os pontos. Nela a distância entre os pontos é calculada em blocos, em alusão à organização da cidade de *Manhattan*, que para se chegar a uma determinada avenida é necessário percorrer os blocos entre os prédios, não sendo possível percorrer uma linha reta.

O cálculo para distância *Manhattan* é apresentado na Equação 2.3.

$$dist(E_i, E_j) = \sum_{l=1}^{M} |x_{il} - x_{jl}|$$

Equação 2.3 - Distância Manhatan

2.3.3.4 - Coseno

Salton (1983) afirma que a distância do coseno é uma medida clássica utilizada na busca de informações sendo uma medida amplamente usada para calcular a similaridade de vetores.

A Equação 2.4 apresenta como é calculada a distância do coseno.

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{(\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)}\sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)})}$$

Equação 2.4 - Distância do Coseno

Segundo Owen et al (2012) a distância do coseno desconsidera o tamanho dos vetores. Isso pode funcionar corretamente para alguns conjuntos de dados, mas irá gerar clusters de baixa qualidade quando a distância relativa dos vetores for pequena.

2.3.3.5 - Tanimoto

Ainda segundo Owen et al (2012) a distância *Tanimoto* captura a informação sobre o ângulo e a distância relativa entre os pontos. Por exemplo, considerando três vetores A (1.0, 1.0), B (3.0, 3.0), e C (3.5, 3.5), pelos vetores apontarem na mesma direção a distância do coseno é a mesma para os três e ela não reconhece que os vetores B e C são muito próximos. A distância euclidiana refletiria essa pequena diferença, mas não considera o ângulo entre eles. Já a distância *Tanimoto* supre, para esse caso, as falhas geradas pelas distâncias do coseno e euclidiana considerando o ângulo entre os vetores e a distância entre os pontos.

A fórmula para se calcular a distância *Tanimoto* é apresentada na Equação 2.5.

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)} + \sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)} - (a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}$$

Equação 2.5 - Distância *Tanimoto*

2.4 - BASES DE DADOS ACADÊMICAS

As bases de dados acadêmicas são uma ferramenta muito importante para auxiliar pesquisadores no processo de criação de seus trabalhos científicos fornecendo informações sobre as publicações e seus autores.

Essas bases de dados podem ser usadas para pesquisar trabalhos em determinada área de interesse, identificar autores que publicam constantemente numa certa linha de pesquisa, entre outros.

Esta seção tem como objetivo apresentar uma série de publicações que tratam dessas bases de dados científicas, com o intuito de ilustrar o estado da arte atual das pesquisas relacionadas ao tema em questão.

A Tabela 2.1 apresenta a listagem com os principais sistemas bibliométricos e os recursos disponíveis em cada um deles. Em alguns deles o usuário tem a informação explícita sobre o número de citações pelo autor e publicação, em outros esse número pode ser descoberto pela contagem manual da lista em tela.

Tabela 2.1 - Matriz de recursos dos principais sistemas bibliométricos, Março de 2016.

	ACM Portal	CiteSeerX	DBLP	Google Scholar	Scopus	Web of Science
Livre	Parcialmente	Sim	Sim	Sim	Não	Não
Automatizada	Não	Sim	Não	Sim	Não	Não
Número de registros	2.0+ mi.	32.23 mi.	3.27 mi.	Não Disponível	Não Disponível	90 mi.
Toda biblioteca disponível pra download	Não	Sim	Sim	Não	Não	Não
Lista de Referências	Sim	Sim	Parcialmente	Não	Sim	Sim
Lista de Citações	Sim	Sim	Parcialmente	Sim	Sim	Sim
Número de citações para a publicação	Sim	Sim	Parcialmente	Sim	Sim	Sim
Número de citações para o autor	Sim	Indiretamente	Parcialmente Indiretamente	Indiretamente	Sim	Sim
Áreas de domínio	Ciência da Computação	Ciência da Computação	Ciência da Computação	Geral	Geral	Geral

Fiala (2010) faz em sua pesquisa uma extensiva mineração na base de dados do CiteSeerX. Ele ressalta que essa fonte tem sido pouco explorada na literatura científica e que um dos motivos para a baixa exploração dessa base é o receio de que os dados obtidos

de forma automática da internet são imprecisos, incompletos, ambíguos, redundantes ou simplesmente errados.

No entanto as informações contidas no CiteSeerX podem ser mineradas através de web crawler, de arquivos XML que podem ser baixados via Internet, ou através de uma API desenvolvida para facilitar o processo de busca das informações. Essa API pode ser usada tanto para obter toda a base de informações ou também ser configurada com filtros de pesquisa para, por exemplo, possibilitar a obtenção de publicações de uma determinada universidade ou assunto.

Nesse caso, as fontes de informações utilizadas foram os arquivos XML com o conteúdo requerido. E a pesquisa foi feita explorando as informações referentes às citações entre autores com o objetivo de formar rankings dos pesquisadores mais influentes.

O conceito usado para criar os rankings de citação leva em consideração o nível de afinidade entre os autores, por exemplo, uma citação de um colega de universidade e/ou departamento tem menor valor do que uma citação de um pesquisador de uma universidade diferente. A finalidade desse conceito é penalizar autores citados pela frequência na colaboração com os autores que os citam (Fiala 2010).

O trabalho é concluído com uma série de rankings com os autores mais citados e com maiores números de registros, no entanto não considera as filiações dos autores tampouco os assuntos dos quais as publicações tratam.

Na seguinte pesquisa (Zaiane et al 2007) foram usadas as informações contidas na base de dados do DBLP. O objetivo foi a descoberta de conhecimento sobre a comunidade acadêmica e também sobre as colaborações entre autores.

Essa base de dados, assim como a base de dados do CiteSeerX, disponibiliza arquivos XML com os seus dados para download. No entanto, a estrutura original disponibilizada pelo DBLP não abrange o escopo da pesquisa.

Portanto foi necessário minerar os dados referentes às palavras-chaves (keywords) dos artigos, mudando assim a estrutura original e criando a estrutura adicional que relaciona os artigos às suas palavras-chaves, que é apresentada na Figura 2.8.

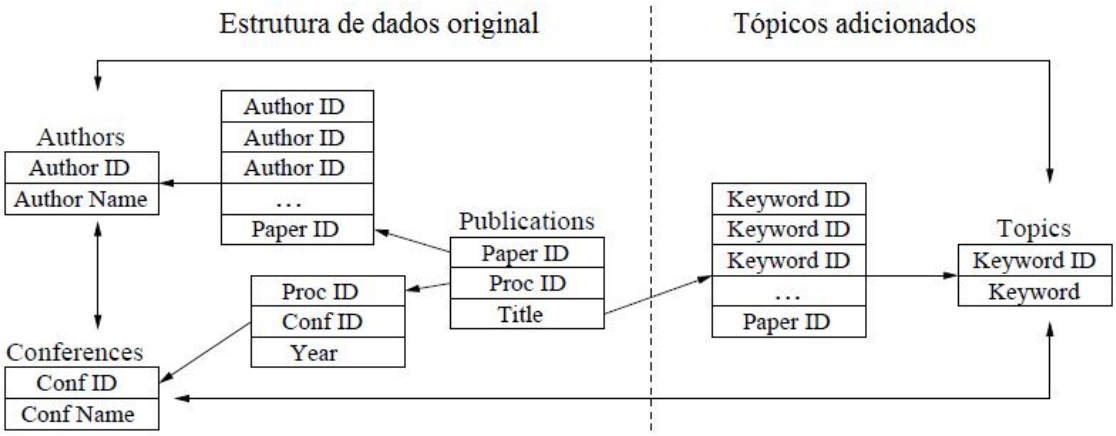

Figura 2.8 - Estrutura de dados extraída do DBLP (Zaiane et al 2007).

No entanto, os valores atribuídos para as palavras-chaves nessa estrutura não são exatamente as mesmas das publicações, pelo fato de que o DBLP não disponibiliza essas informações.

Dessa forma, para atribuir as palavras-chaves às publicações foi usado um método que as extrai a partir do título baseado na frequência das palavras em todos os títulos. Esse método primeiramente remove as palavras de parada (*stopwords*) dos títulos e depois conta a frequência das palavras restantes, contando também as palavras em pares a fim de descobrir os tópicos compostos, por exemplo, Mineração de Dados, Sistemas Distribuídos e assim por diante (Zaiane et al 2007).

O método proposto na pesquisa cria dois tipos de modelos gráficos, o primeiro com os autores e as conferências, e o segundo adicionando os tópicos (palavras-chaves) com o objetivo de descobrir as comunidades no âmbito das conferências, tópicos e autores.

Como resultado foi criado uma aplicação com uma interface interativa em que o usuário pode navegar a partir de um tópico, conferência ou autor e visualizar as listagens com as comunidades.

Por exemplo, a partir de um tópico é apresentada a listagem com as conferências e pesquisadores relacionados, a partir de uma conferência são apresentados os pesquisadores, tópicos e as conferências relacionadas e a partir de um pesquisador são apresentadas as conferências, os tópicos, e os pesquisadores relacionados, assim como os coautores e as colaborações.

As listas de conferências relacionadas, tópicos e pesquisadores para um autor podem ser usadas para entender as entidades próximas e as comunidades de pesquisa. (Zaiane et al 2007).

Na próxima pesquisa os autores (Zhang et al 2010) usam a base de dados 'Web of Science of Thomson-Reuters' para fazer um estudo focado na análise do fluxo das informações entre os assuntos e categorias e encontrar uma estrutura usando os assuntos das publicações para criação dos clusters.

Como resultados eles criaram uma série de rankings com as categorias das publicações, como por exemplo, qual categoria possui mais de 150 publicações em revistas, as 10 primeiras com maior número de entropias, as 10 primeiras categorias na área de ciência e ciências sociais, entre outros.

Por fim, foram gerados mapas com as citações cruzadas entre as categorias nas publicações.

No próximo trabalho os autores (Leydesdorff and Bihui 2005) usaram a base de dados 'Science Citation Index (SCI;Thomson ISI, 2002)' e a base de dados 'Chinese Science Citation' da Academia Chinesa de Ciências para descobrir as estruturas de citação entre as publicações usando as citações entre revistas.

Eles também mapearam a estrutura de citações considerando o tipo de evento das publicações. Por exemplo, foi encontrado um artigo de revista da engenharia da computação fazendo citação a um artigo de uma revista de química.

Outra abordagem pode ser encontrada no trabalho dos autores (Choi et al 2013) no qual a análise é feita baseada na base de dados de citações da Korea (Korea Science Citation Database (KSCD)).

Nessa pesquisa foi feita a comparação entre publicações que não foram escritas em inglês, vindas das regiões da Ásia e do Pacífico, com publicações originárias de bases de dados que têm o inglês como língua principal para o seu conteúdo, por exemplo, Reuter's Web of Science, Elsevier's Scopus e Google Scholar.

Os autores também incluíram as informações provindas da base de dados KSCD, que é uma base de dados digital para os trabalhos produzidos na Coreia.

Nesse trabalho foram analisados a evolução e o impacto dos artigos Coreanos em relação aos trabalhos publicados em inglês, dividindo em categorias, área de atuação, assunto, tipo de publicação e referências.

Por fim, como resultados foram apresentados mapas de clusterização da colaboração entre as universidades Coreanas.

Em sua pesquisa Sinoara et al (2014) propõe a utilização de entidades nominais para a criação de clusters hierárquicos de forma incremental. Entidades nominais são aquelas que podem ser identificadas por um nome próprio, por exemplo, pessoas, organizações, localizações, marcas e produtos.

Para analisar o impacto das entidades nominais nos clusters hierárquicos textuais eles fazem um experimento utilizando três bases de dados, sendo a primeira uma base de dados sobre artigos esportivos escritos em Português, a segunda um subconjunto de 20 grupos de notícias e a terceira a base de dados Reuters que contém artigos científicos escritos em inglês.

Segundo Sinoara et al (2014) o modelo bolsa de palavras (*bag-of-words*) é o mais popular para representação de clusters de textos. Nesse modelo as características das palavras dos documentos são usadas onde a semântica e a ordem das palavras não são consideradas. Essa representação pode ser tomada como informação técnica dos dados, que é obtida através de uma tarefa simples da extração das palavras dos documentos, com a geração de um baixo custo computacional.

Eles justificam o uso das entidades nominais para geração de informações e clusters por três razões principais:

1. Entidades nominais representam informações ricas sobre o conteúdo dos documentos, e isso é potencialmente útil para complementar e refinar o conhecimento extraído das informações técnicas;
2. Uma das características principais da existência de informações privilegiadas é que ela está disponível apenas para uma fração dos documentos. Essa característica também é natural do reconhecimento das entidades nominais, pois algumas coleções de documentos podem ter uma fração ou toda a coleção de entidades nominais não reconhecidas;

3. Entidades nominais não estão explicitamente disponíveis nos dados e seu correto reconhecimento requer algum processamento adicional que normalmente usa algum tipo de inteligência e requer um alto custo computacional. Esse fato faz com que o reconhecimento de entidades nominais em grandes coleções de dados seja impraticável. No entanto pode ser feito para uma amostra de documentos, e dessa forma as entidades nominais podem ser usadas para obtenção de informações privilegiadas. (Sinoara et al 2014).

Com o uso dessas técnicas eles afirmam conseguir clusters com maior precisão que os clusters gerados de acordo com as técnicas tradicionais existentes, e que esses clusters gerados a partir de entidades nominais além de serem mais precisos são mais entendíveis para os usuários e não necessitam de nenhum processamento adicional para a descoberta dos centroides.

A pesquisa de Benghabrit et al (2013) faz o uso da base de dados Reuters como entrada de dados e propõe uma técnica que reúne termos estatísticos e semânticos relevantes para guiar os mecanismos de criação dos clusters a partir de fontes de dados não estruturadas.

Eles afirmam que no mecanismo de clusterização de textos, um documento é tipicamente representado por um modelo de vetores esparsos. Consequentemente, o vocabulário consiste nos termos únicos contidos nos documentos, o que significa a geração de uma quantidade massiva de dados.

Com o modelo vetorial e uma grande massa de dados, os algoritmos de clusters se tornam bastante complexos e menos precisos. De fato, as distâncias entre cada par de documentos se tornam quase as mesmas para a maioria das distâncias de medida.

Outro problema é que algumas dimensões e atributos relacionados ao cluster são irrelevantes ou redundantes, o que pode influenciar o processo de clusterização de forma negativa.

Como consequência é fortemente recomendado diminuir a massa de dados a ser analisada selecionando os dados mais representativos antes de realizar o processo de clusterização. Dessa forma é possível diminuir a complexidade dos métodos de clusterização e aperfeiçoar a eficiência na criação dos clusters.

Assim a seleção das características relevantes é o método mais adequado para redução do tamanho do vocabulário, obtenção de melhor entendimento dos dados e a geração de clusters mais precisos e conclusivos.

Nesse cenário, de uma grande massa de dados de documentos textuais e com o uso da técnica que combina simultaneamente as análises estatísticas e semânticas, eles afirmam conseguir clusters com maior precisão que os clusters gerados de acordo com as técnicas tradicionais existentes que são geradas sem a combinação das análises estatísticas e semânticas, usando puramente apenas uma delas para obtenção dos resultados.

Chikhaoui et al (2015) aborda a evolução da influência das citações nas redes sociais dinâmicas e usa a base de dados do DBLP como estudo de caso para sua pesquisa.

Nessa pesquisa é focada a evolução da influencia nas citações entre as comunidades nas redes sociais. Para tanto é usado um multidiagrama gráfico temporal para representar a dinâmica da rede social e analisar a influência das relações entre as comunidades ao longo tempo.

Thasleena et al (2014) propõe um método para classificar documentos XML eliminando predições de classes ambíguas. Em seu método são contidos vários passos para o processamento dos arquivos, que são:

1. **Fase de aprendizado: Pré-processamento**
    - Remove os *stop words*;
    - Etapa de tokenizazação onde os documentos XML são processados para gerar *tokens* a partir das *tags* e atributos dos arquivos;
    - Os termos remanescentes no conteúdo são transformados nas suas respectivas formas originais;
    - Divide os arquivos XML na base de suas classes.
2. **Fase de aprendizado: Extração de características**

O caminho do elemento principal de cada termo no conteúdo do documento XML se torna uma chave única onde o nó folha será o termo. A lista de chaves únicas recuperadas a partir de cada documento XML pertencente a mesma classe são armazenados no mesmo grupo.

3. **Fase de Aprendizado: Regras de mineração**

Algoritmos de aprendizado são aplicados na lista de arquivos com a mesma classe para recuperar a frequência das chaves.

4. **Fase de Aprendizado: Seleção dos dados**

Regras são geradas a partir dos termos com maior frequência. As regras geradas serão exploradas pela classificação associativa para os documentos XML.

5. **Fase de Teste**

Os documentos XML na lista de teste devem ser classificados pela classificação associativa.

Como resultado eles afirmam que a metodologia proposta para classificação dos arquivos XML obteve 100% de precisão ao classificar a base de dado do DBLP.

Tchuente et al (2013) afirma que atualmente as redes sociais são amplamente usadas como solução para enriquecimento da informações sobre perfis de usuários como, sistemas de recomendação ou sistemas personalizados. Essas informações de perfis podem ser uma fonte significativa para geração de redes de interesse.

No entanto eles afirmam que as técnicas existentes dão foco individual no usuário da rede social o que gera uma forte dependência sobre o objetivo de cada autor.

Para melhorar essas técnicas eles propõem um algoritmo baseado em comunidades que é aplicado em parte da rede social dos usuários e que deriva dos perfis dos usuários na rede social que podem ser usados para qualquer propósito.

Foram computados os interesses dos usuários por essas comunidades considerando a semântica (interesses relacionados às comunidades) e as suas medidas estruturais (por exemplo, medidas de centralidade dos nós da rede) para geração de um gráfico de redes de relacionamento.

Para gerar os resultados do experimento foram usados dados recuperados do Facebook e do DBLP para aplicação do algoritmo proposto e geração dos resultados com o cruzamento de dados entre as fontes de informação.

Way et al (2016) faz um estudo sobre a influência dos autores na área da ciência da computação separando-os por gênero (masculino e feminino) e conclui que as mulheres representam apenas 20% dos pesquisadores com grau de doutorado, sendo 18% delas com

solteiras, e estimasse que apenas 20% dos cargos técnicos da industria da computação sejam ocupados por mulheres.

Na área acadêmica a tendência é a mesma sendo que as mulheres representam apenas 15% dos cargos de professores efetivos nos departamentos de ciência da computação das universidades pesquisadas.

Em seu estudo Way et al (2016) tenta compreender as causas desse desequilíbrio de gênero no que se refere a representação das mulheres na área de ciência da computação, seja na área acadêmica ou profissional. E com a ilustração das prováveis causas dessa desigualdade tentar prover solução para apoiar a igualdade de oportunidades independentemente do gênero do candidato.

Blondel et al (2008) propõe um método para extrair comunidades a partir de grandes redes de dados. Eles apresentam um método heurístico baseado na otimização da modularidade, e também como realizar métodos de detecção de comunidades em tempo real.

O problema da detecção de comunidades é que ela exige que a rede seja dividida em comunidades em que os nós são densamente conectados, e que pertençam a outras comunidades sejam escassamente conectados.

Implementações precisas desse problema de otimização são conhecidas como computacionalmente intratáveis. Portanto, vários algoritmos têm sido propostos para achar partições razoavelmente precisas com um curto tempo de processamento.

Para tanto, Blondel et al (2008) introduz um algoritmo que encontra partições de grandes redes em um curto período de tempo e que desdobra toda a estrutura hierárquica de comunidades dessa rede. Assim, dando acesso a diferentes resoluções de detecção de comunidades.

No estudo de caso Blondel et al (2008) usam a rede belga de telefonia móvel para aplicação do algoritmo criado para encontrar comunidades. E com o uso desse algoritmo foi possível executar o processo de identificação das comunidades dentro da rede em questão.

2.3 - SÍNTESE DO CAPÍTULO

Este capítulo teve como objetivo apresentar os principais conceitos, tecnologias e os trabalhos relacionados ao que foi desenvolvido nesta pesquisa. Foram tratados os conceitos de extração transformação e carga, mineração de dados, clusters tais como os tipos de clusters, algoritmos e medidas de distâncias.

Também foram trazidos os trabalhos relacionados a essa pesquisa, que tratam de mineração de dados em bases acadêmicas, clusterização tanto de fontes de dados estruturadas e não estruturadas, criações de rankings e comunidades de relacionamento.

3. PROPOSTA DE MODELO PARA CLUSTERIZAÇÃO

Este capítulo visa apresentar o modelo proposto nesta pesquisa, englobando as etapas necessárias para alcançar os objetivos estabelecidos neste trabalho, que são o processo de extração transformação e carga, a mineração dos dados adicionais, tratamento e normalização dos dados, preparação para criação dos clusters, criação dos rankings e redes de relacionamento.

O modelo proposto foi criado visando ser aplicável a qualquer base de dados acadêmica, mas para fins de validação e para criação do estudo de caso foram usadas as bases de dados do CiteSeerX e do DBLP, que foram escolhidas por serem bases que têm como foco trabalhos nas áreas de ciência da computação e afins.

3.1 – DIAGRAMA ARQUITETURAL

Para possibilitar o alcance dos objetivos deste trabalho foram criados vários artefatos de software que em conjunto compõem um modelo arquitetural englobando todo o processo executado neste trabalho para possibilitar a geração dos resultados e o alcance dos objetivos.

A Figura 3.1 apresenta o diagrama arquitetural que contempla todos os componentes criados para possibilitar o alcance dos objetivos deste trabalho e os organiza em forma de processo desde a extração dos dados até a criação dos resultados.

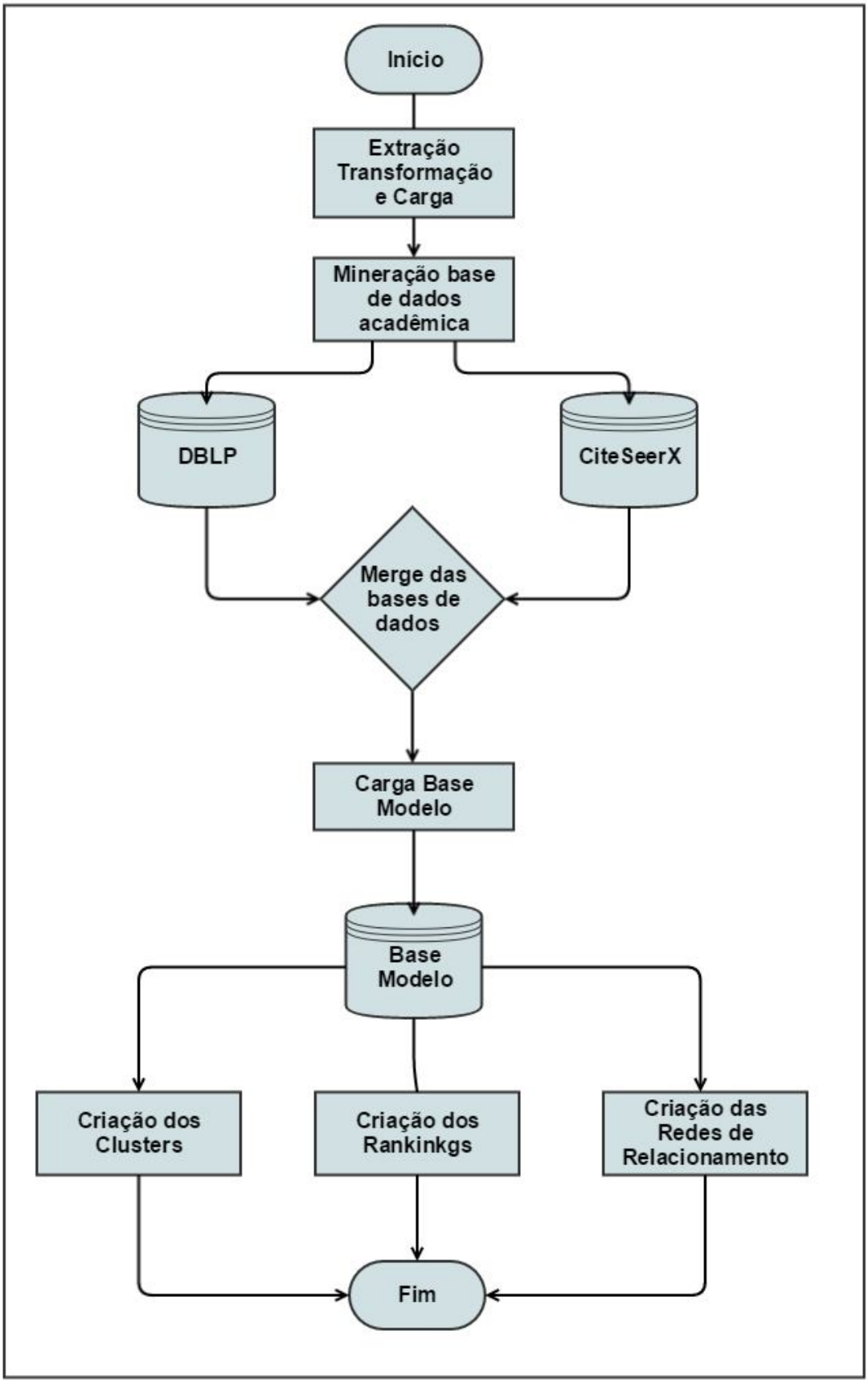

Figura 3.1- Diagrama arquitetural

O processo obedece aos seguintes passos principais:

- Extração transformação e carga;
- Mineração de dados (dados adicionais);
- Carga da Base Modelo;
- Criação dos Resultados.

Nas seções a seguir os passos do processo serão explicados detalhadamente conforme a estrutura apresentada abaixo:

- Extração transformação e carga
    - Fonte de dados;
    - API de mineração;
    - Escopo mineração;
    - Carga.
- Mineração (dados adicionais)
    - Escopo mineração;
    - Dados adicionais;
    - Algoritmos de mineração (web crawlers);
        - CiteSeerX;
        - DBLP.
- Bases de dados mineradas;
- Normalização das bases de dados;
    - Base de dados genérica.
- Carga na Base Modelo;
    - ETL.
- Geração dos resultados;
    - Aplicação dos processos para gerar os resultados.

3.1 - EXTRAÇÃO TRANSFORMAÇÃO E CARGA

A etapa inicial do estudo de caso desta pesquisa foi a extração transformação e carga (ETL) dos dados necessários para criação dos clusters, dos rankings e das redes de comunidade de relacionamento.

Nessa etapa de ETL foram usadas formas distintas para obtenção dos dados, a primeira foi a extração dos dados do CiteSeerX usando a API provida e a segunda foi a extração dos dados do DBLP usando os arquivos XML disponibilizados pela ferramenta.

3.1.1 - EXTRAÇÃO DOS DADOS DO CITESEERX

O CiteSeerX provê uma API para extração dos dados chamada OAI-Harvest. Essa API possibilita que os dados sejam visualizados ou baixados programaticamente via URL. Também é possível criar uma aplicação que faça o download dos dados do CiteSeerX usando a API.

Segundo Wu (2014) técnicas de inteligência artificial são usadas em vários componentes do CiteSeerX, incluindo a classificação de documentos, duplicações de autores, gráfico de citações, extração automática de metadados, desambiguação de autores e assim por diante.

Para extrair os dados utilizados nesta pesquisa foi desenvolvida uma aplicação usando a API disponibilizada pelo CiteSeerX. Nessa aplicação foi criada uma série de algoritmos usados para processar os dados obtidos. Esses algoritmos serão detalhados nas seções a seguir.

3.1.2 - MÉTODO UTILIZADO CITESEERX

A Figura 3.2 ilustra os passos do algoritmo de extração de dados que foi concebido e otimizado para processar os dados provenientes do CiteSeerX através da API provida.

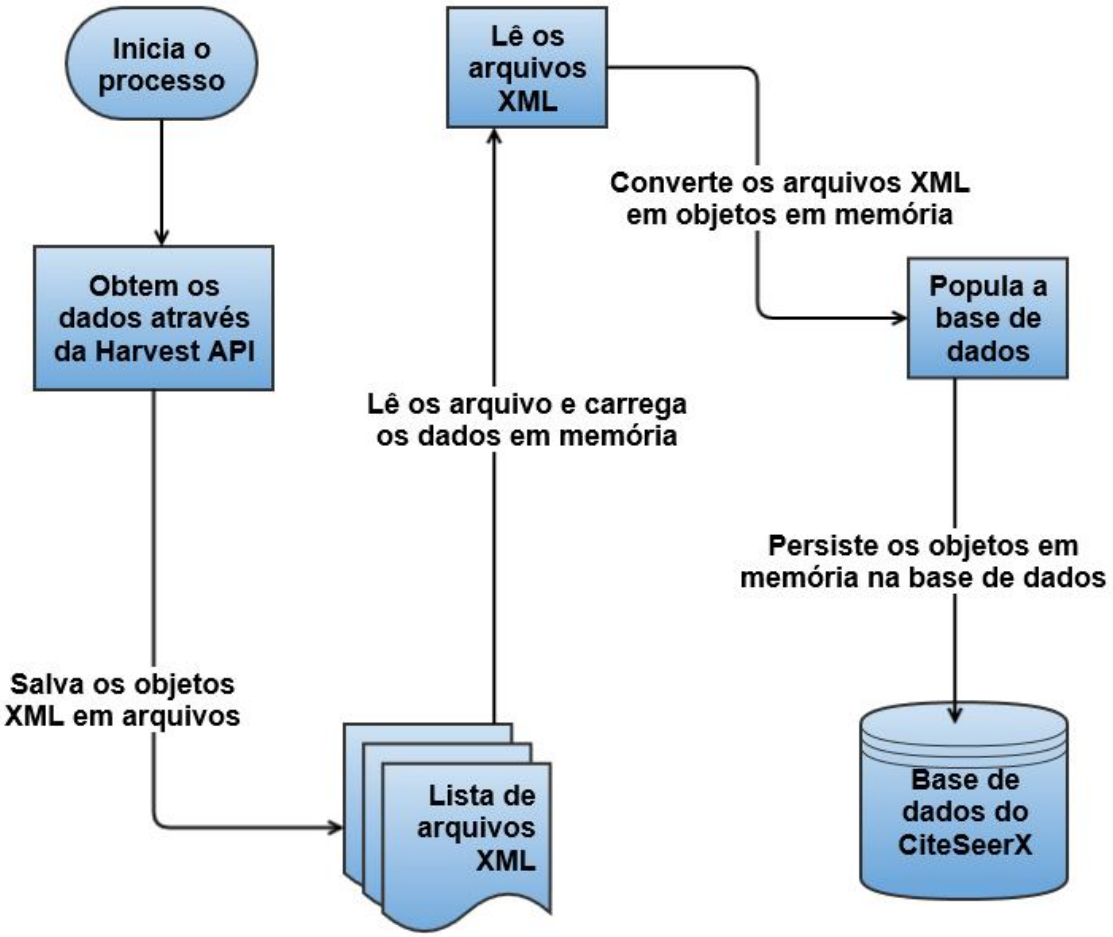

Figura 3.2 - Passos do algoritmo de mineração.

O funcionamento ocorre seguindo os passos descritos abaixo:

1. Para cada repetição são retornados um bloco com 25000 registros (valor padrão fornecido pela API) pela API do CiteSeerX;
2. É gravado um arquivo XML em disco com os 25000 registros retornados;
3. Os passos 1 e 2 são repetidos até todos os dados requisitados serem retornados;
4. Faz-se a leitura de cada arquivo gerado nos passos anteriores;
5. Os dados do arquivo em memória são carregados;
6. O arquivo XML é convertido em objetos;
7. Os objetos perduram no banco de dados;
8. Repetem-se os passos 4 a 7 até que todos os dados de todos os arquivos gerados sejam salvos no banco de dados.

3.1.3 - ESTRUTURA DE DADOS DO CITESEERX

A Figura 3.3 apresenta a estrutura do XML provido pela OAI-Harvest API. A estrutura contém os campos referentes ao cabeçalho e aos metadados dos registros. O bloco do cabeçalho contém as informações de data e identificador do registro. O identificador pode ser usado via URL para acessar o registro no site do CiteSeerX.

Nos metadados são encontrados os campos usados para a execução deste estudo de caso, são eles:

- *dc:title*. Título da publicação;
- *dc:creator*. Lista dos autores da publicação;
- *dc:subject*. Lista com os assuntos de que a publicação trata;
- *dc:description*. Descrição da publicação;
- *dc:contributor*.Colaborador;
- *dc:publisher*. Publicador;
- *dc:date*. Lista com as datas da publicação.
- *dc:format*. Formato do arquivo fonte;
- *dc:type*. Tipo do arquivo fonte (texto, imagem, etc);
- *dc:identifier*. URL para acesso via navegador de Internet;
- *dc:source*. URL para acesso ao arquivo fonte da publicação;
- *dc:language*. Idioma da publicação;
- *dc:rights*. Informações referentes aos direitos autorais da publicação.

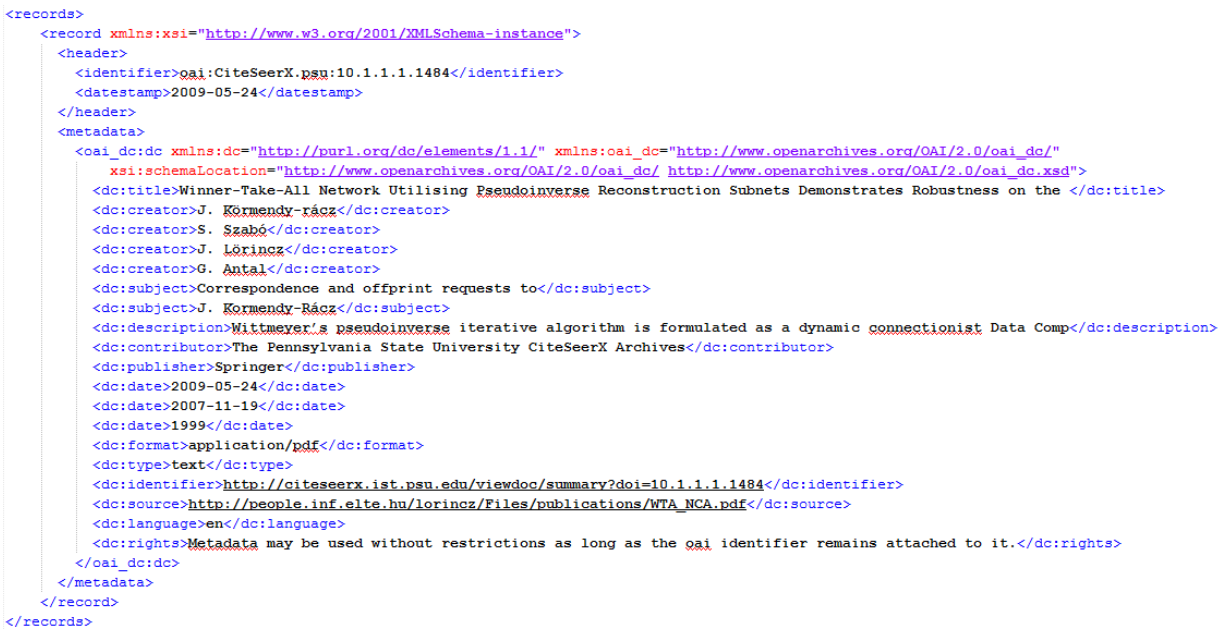

Figura 3.3 - Estrutura do XML usada na mineração através da API do CiteSeerX

3.1.4 - MODELO DE ENTIDADE E RELACIONAMENTO CITESEERX

A partir da estrutura disponibilizada no XML foi concebido um modelo de entidade e relacionamento (MER) para criação da base de dados usada nesta pesquisa. Inicialmente a base de dados foi criada com o objetivo de reproduzir uma cópia fiel da estrutura disponibilizada pela API de mineração.

A Figura 3.4 apresenta o MER criado. Nele a tabela '*document*' armazena as publicações e se relaciona com as tabelas que armazenam os assuntos e os autores, nas quais uma publicação pode ter um ou mais assuntos e autores.

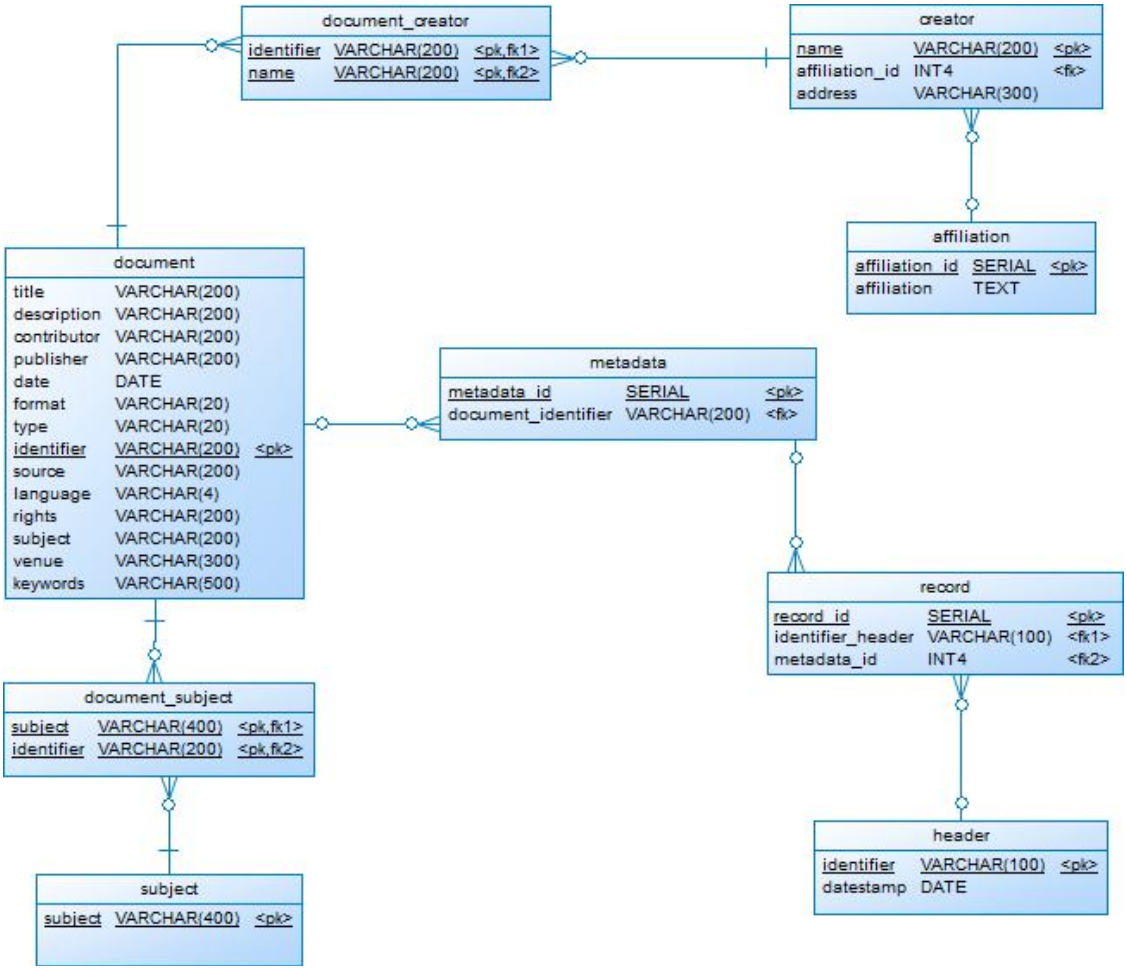

Figura 3.4 - Modelo de entidade e relacionamento do CiteSeerX

3.1.5 - EXTRAÇÃO DOS DADOS DO DBLP

O DBLP é uma biblioteca online aberta focada em trabalhos da área de ciência da computação. Ela foi criada por um grupo de pesquisa da universidade de Trier, na Alemanha, com o foco em publicações nas áreas de sistemas de banco de dados e lógica de programação, e posteriormente foi sendo expandida gradualmente para todas as áreas da ciência da computação.

Os dados providos pelo DBLP são disponibilizados em uma série de formatos, como páginas de internet e arquivos em formato XML, JSON, RDF e BibTeX.

3.1.6 - MÉTODO UTILIZADO DBLP

O DBLP disponibiliza para download vários arquivos XML com o conteúdo de sua base de dados. Tais arquivos podem ser recuperados programaticamente ou manualmente via requisições HTTP.

Para processamento dos arquivos XML o DBLP provê uma série de fontes escritas em Java que leem o conteúdo dos arquivos e os converte para objetos em memória.

A Figura 3.5 apresenta os passos percorridos pelo algoritmo criado para extrair os dados do DBLP.

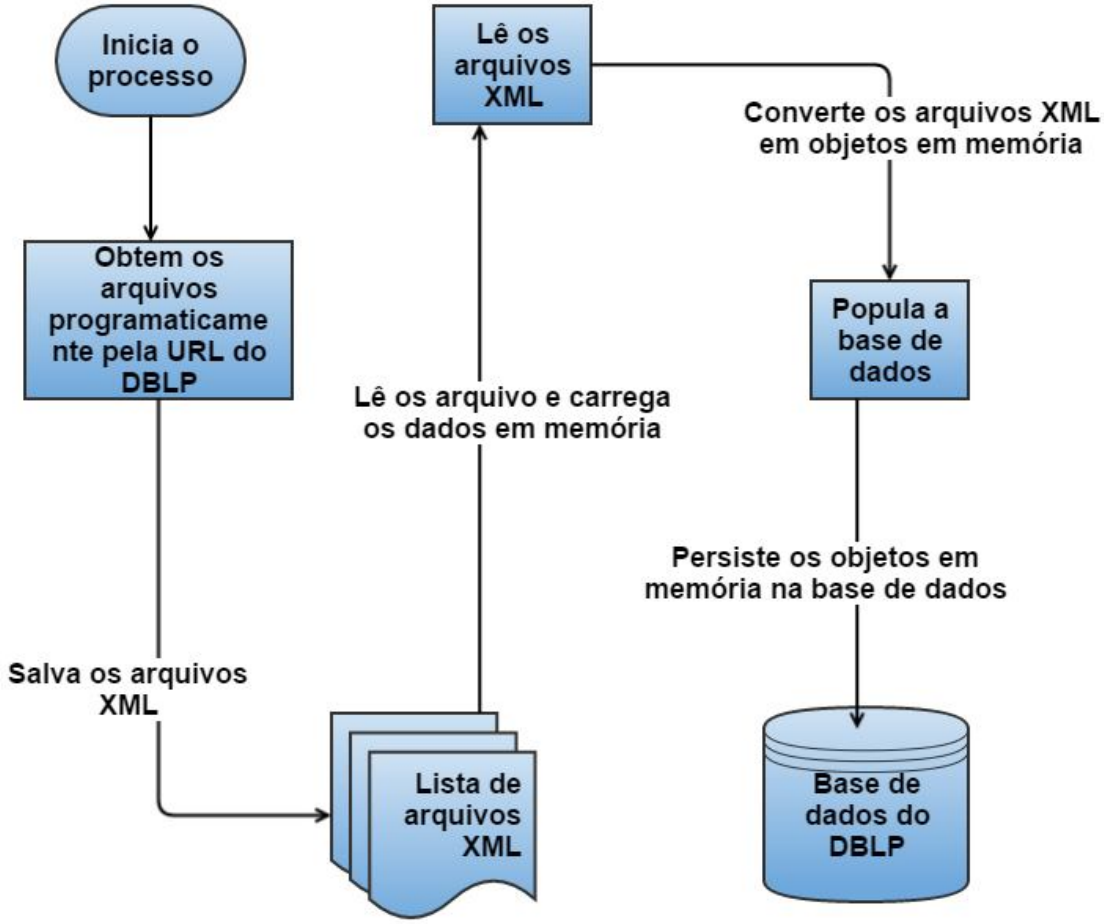

Figura 3.5 - Passos do algoritimo de extração do DBLP

O funcionamento do algoritmo obedece aos seguintes passos:

1. Para cada repetição é obtido um arquivo XML com os dados;
2. O arquivo XML é salvo em disco;
3. Os passos 1 e 2 são repetidos até que todos os dados requisitados sejam retornados;

4. Faz-se a leitura de cada arquivo gerado nos passos anteriores;
5. Carregam-se os dados do arquivo em memória;
6. O arquivo XML é convertido em objetos.
7. Os objetos perduram no banco de dados
8. Repetem-se os passos 4 a 7 até que todos os dados de todos arquivos gerados sejam salvos no banco de dados.

3.1.7 - ESTRUTURA DE DADOS DO DBLP

O DBLP disponibiliza que seus dados sejam extraídos em vários formatos, tais como JSON, RDF, XML e BibTeX. Para os fins desta pesquisa foi usado o formato provido em XML para a extração dos dados.

A Figura 3.6 apresenta a estrutura do arquivo XML provido pelo DBLP para extração dos seus dados. Diante dessa estrutura foi concebido o modelo de entidade e relacionamento que foi usado como artefato para obtenção dos resultados.

O DBLP tem como fonte de informação oito categoria de dados, que são:

- *Article*;
- *Inproceedings*;
- *Proceedings*;
- *Book*;
- *Incollection*;
- *Phdthesis*;
- *Mastersthesis*;
- *WWW*.

Para cada uma das categorias citadas acima o DBLP disponibiliza a mesma estrutura de dados que é exemplificada na Figura 3.6.

```
<articles>
    <article key="journals/jods/HurtadoPW08" mdate="2008-04-15">
        <author>Carlos A. Hurtado</author>
        <author>Alexandra Poulovassilis</author>
        <author>Peter T. Wood</author>
        <title>Query Relaxation in RDF.</title>
        <pages>31-61</pages>
        <year>2008</year>
        <volume>10</volume>
        <journal>J. Data Semantics</journal>
        <ee>http://dx.doi.org/10.1007/978-3-540-77688-8 2</ee>
        <crossref>journals/jods/2008-10</crossref>
        <url>db/journals/jods/jods10.html#HurtadoPW08</url>
    </article>
</articles>
```

Figura 3.6 - Estrutura do XML do DBLP

A seguir são detalhados os atributos do XML disponibilizado pelo DBLP.

- *author*: Nome do autor;
- *editor*: Editor da Publicação;
- *title*: Título da publicação;
- *booktitle*: Título do livro;
- *pages*: Número de páginas;
- *year*: Ano da publicação;
- *address*: Endereço;
- *journal*: Revista da publicação;
- *volume*: Volume;
- *number*: Número;
- *month*: Mês;
- *url*: Endereço eletrônico para a publicação;
- *ee*: Endereço para a referência da publicação;
- *publisher*: Publicador;
- *note*: Notas;
- *crossref*: Referência cruzada;
- *isbn*: Número do ISBN;
- *series*: Número de Série;

- *school*: Instituição;
- *chapter*: Capítulo.

3.1.8 - MODELO DE ENTIDADE E RELACIONAMENTO DBLP

Assim como no processo de extração do CiteSeerX o processo de ETL do DBLP também fez o uso da estrutura disponibilizada no XML para conceber o MER para criação da base de dados do DBLP usada nesta pesquisa.

A Figura 3.7 apresenta o MER do DBLP. Nele é possível observar que as oito tabelas com a maior quantidade de colunas representam as categorias de dados armazenadas, e as tabelas menores localizadas no centro do modelo representam os relacionamentos entre os autores e as publicações.

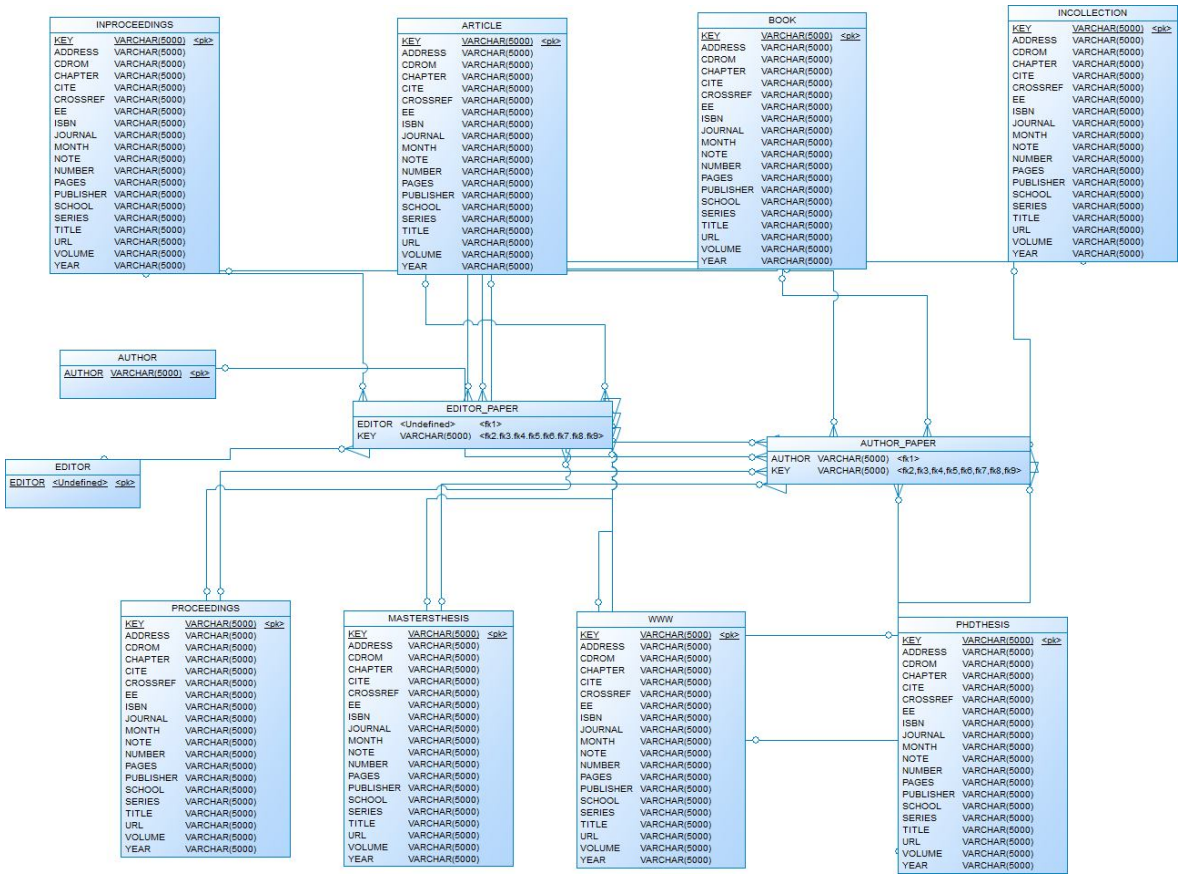

Figura 3.7 - Modelo de entidade e relacionamento do DBLP

3.2- MINERAÇÃO DE DADOS (DADOS ADICIONAIS)

No entanto, alguns campos necessários para a execução deste trabalho de pesquisa não são disponibilizados no XML, logo foram obtidos através de uma mineração adicional que fez o uso de *webcrawlers* criados especificamente para a busca desses dados adicionais.

Essa mineração foi feita através de algoritmos criados para obter as informações necessárias para construção do modelo proposto nessa pesquisa. Esses algoritmos fazem parte de um processo que sincroniza a chamada de execução dos algoritmos de mineração de dados.

Cada um dos quatro algoritmos possui uma série de etapas que são executadas sequencialmente até a obtenção e persistência das informações requeridas pelo banco de dados.

A seguir são detalhados os algoritmos criados para mineração dos dados adicionais do CiteSeerX e do DBLP.

3.2.1 MINERAÇÃO CITESEERX

A Figura 3.8 apresenta, em forma de diagrama, os passos que os algoritmos percorrem para a obtenção dos dados adicionais necessários para a obtenção dos resultados.

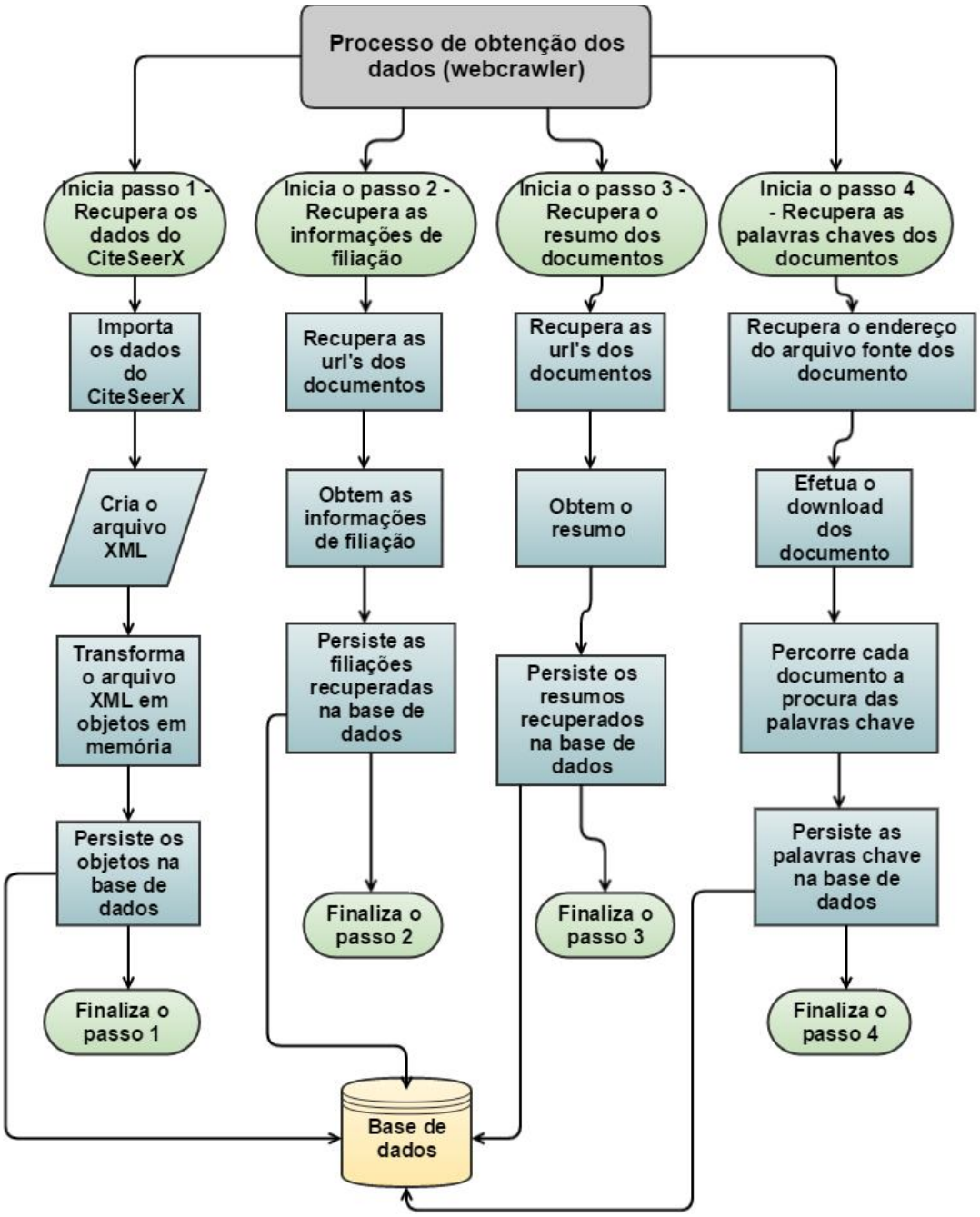

Figura 3.8 - Passos do webcrawler do CiteSeerX

O primeiro algoritmo obtém as informações referentes à filiação dos autores das publicações. As etapas executadas são discriminadas a seguir:

1. Consulta-se o banco de dados, especificamente a tabela *'document'* e retorna os valores do campo *'identifier'* que remete para o endereço da publicação no site do CiteSeerX;

2. Altera-se o valor de cada *'identifier'* obtido de modo que a requisição aponte para a seção de versões do registro. Nessa seção são encontradas as informações referentes à filiação dos autores;
3. O código HTML é obtido da página requisitada e faz-se a busca das filiações dos autores da publicação em questão;
4. Após a obtenção das informações da página HTML, elas são relacionadas com a publicação corrente e persiste a filiação na base de dados associando-as com os respectivos autores;
5. Repetem-se os passos 2 a 4 até que todos os registro retornados pelo passo 1 sejam percorridos pelo algoritmo.

O objetivo do segundo algoritmo é recuperar o resumo das publicações. Assim como o anterior, ele segue uma série de passos para obtenção da informação requerida.

1. Assim como o algoritmo anterior o primeiro passo é recuperar os valores do campo *'identifier'* para os registros da tabela *'document'*;
2. O valor do campo remete diretamente para a seção na qual é encontrado o resumo da publicação. Nesse caso, o algoritmo busca dentro do código HTML obtido pela requisição a seção em que é encontrada a informação referente ao resumo;
3. A informação perdura na base de dados;
4. Repetem-se os passos 2 e 3 até que todos os registros da tabela *'document'* passem pelo processo.

No terceiro e último algoritmo de mineração dos campos adicionais o objetivo é recuperar as palavras-chaves (*keywords*) das publicações. Assim como nos demais há uma série de passos a serem seguidos:

1. Recupera-se o campo source da tabela *document*. Esse campo contém a URL para o arquivo fonte da publicação, que para todos os registros obtidos nessa pesquisa é um arquivo PDF;
2. A partir da URL da publicação faz-se o download do arquivo, que é armazenado em memória;
3. Em tempo de execução abre o arquivo baixado e procura pela palavra *'keyword'* nas cinco primeiras páginas do documento.
4. Se a palavra *'keyword'* for encontrada o que segue imediatamente é armazenado após o sinal de dois pontos;

5. Se a informação for encontrada perdura no banco de dados, senão persiste no campo *keyword* o valor 'não encontrado';
6. Repetem-se os passos 2 a 5 até que sejam percorridos todos os valores retornados pelo passo 1.

Os campos obtidos através dessa segunda mineração são apresentados na Figura 3.9. Nela os campos à esquerda são recuperados através da API provida, já os campos à direita são obtidos pelo processo de mineração descrito, em que os campos *'affiliation'*, *'venue'* e '*venue_type*' são obtidos pelo primeiro algoritmo, e os campos 'abstract' e *keywords'* são recuperados pelos segundo e terceiro algoritmos respectivamente.

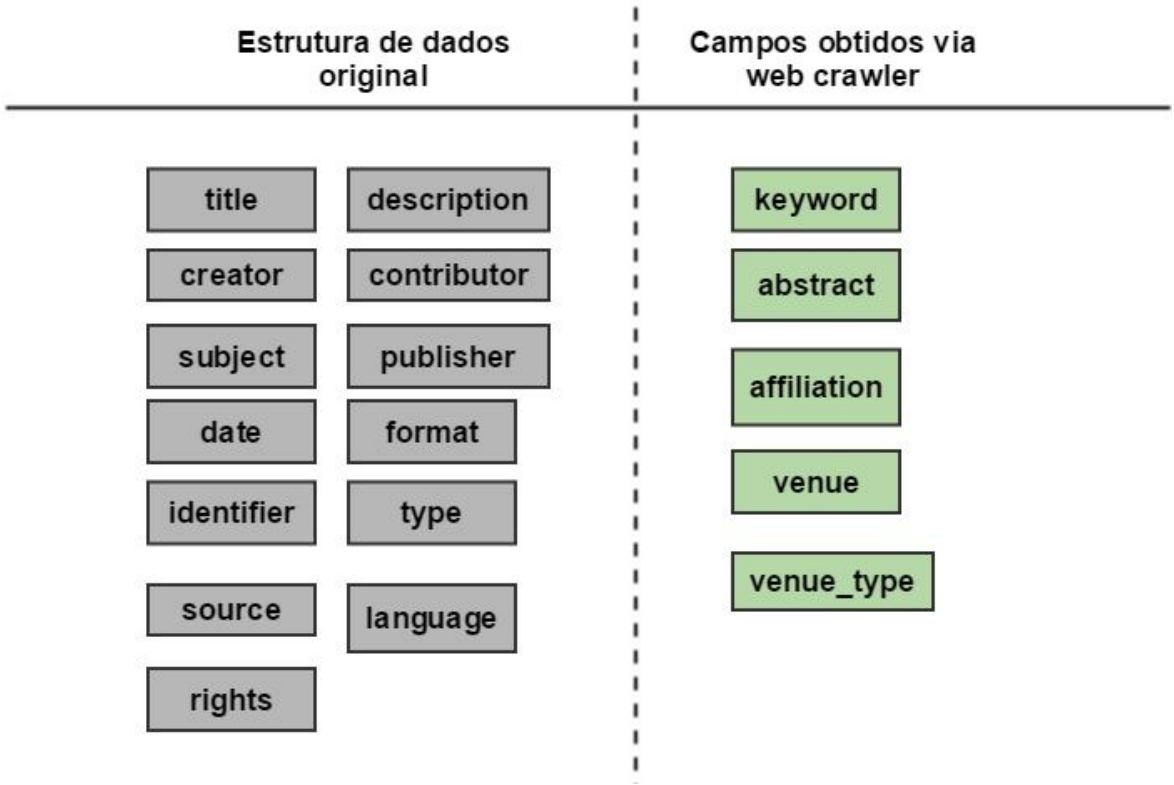

Figura 3.9 - Fontes de dados dos campos adicionais do CiteSeerX

3.2.2 – MINERAÇÃO DBLP

A Figura 3.10 apresenta o diagrama dos passos percorridos pelos algoritmos que extraem os dados adicionais do DBLP necessários para a obtenção dos resultados.

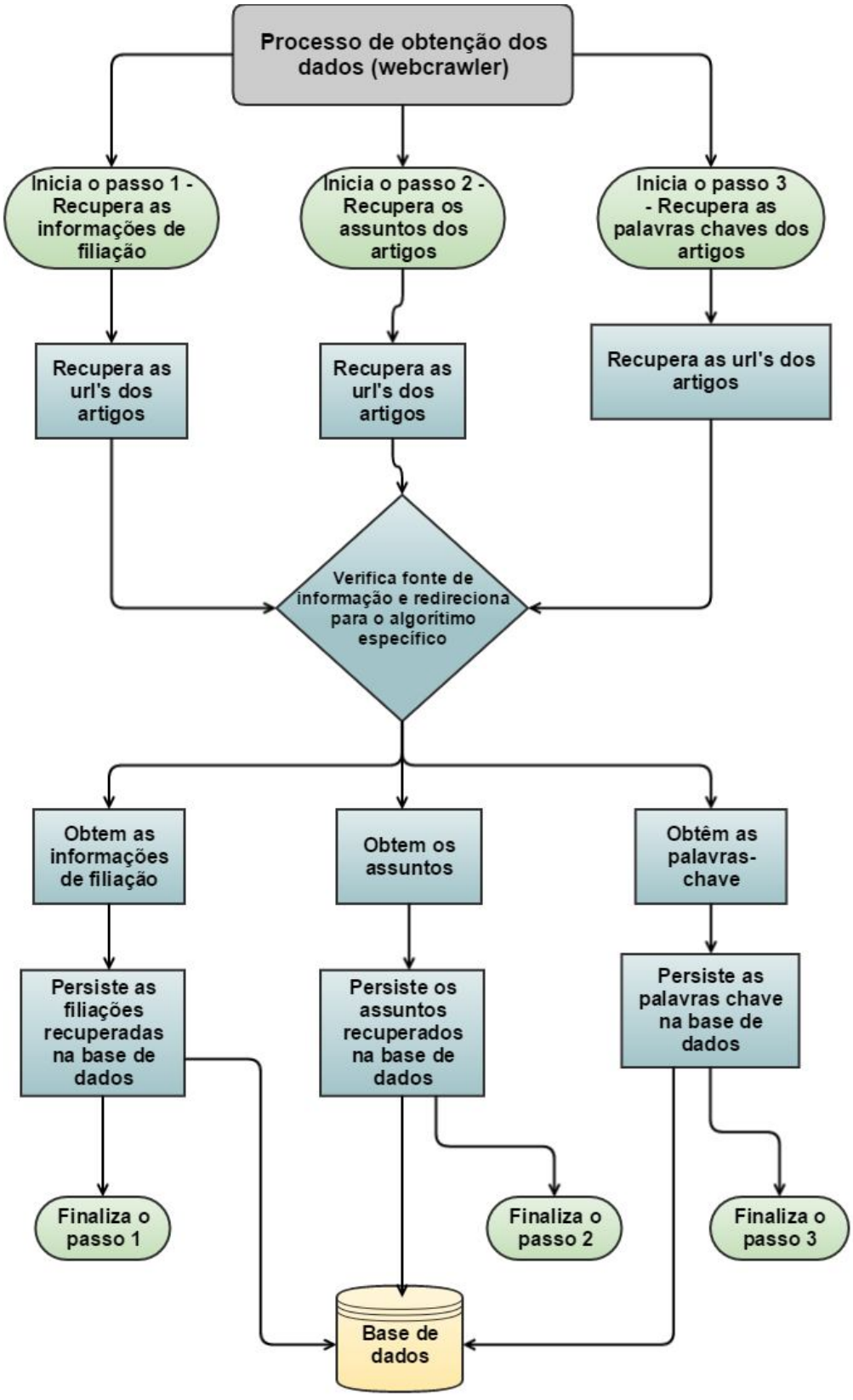

Figura 3.10 - Passos do webcrawler do DBLP

No primeiro algoritmo são obtidas as informações referentes à filiação dos autores das publicações. As etapas executadas pelo programa são apresentadas a seguir:

1. Consulta o banco de dados especificamente à tabela *'article'* e retorna os valores do campo *'ee'* que remete para o endereço da publicação.
2. Verifica para qual site o valor do campo *ee* está direcionando.
3. Identifica o site da publicação e redireciona para o algoritmo especifico de cada *website*.
4. Obtém o código HTML da página requisitada e faz a busca das filiações dos autores da publicação.
5. Após obter as informações da página HTML e relaciona-la com a publicação corrente, persiste a filiação na base de dados associando-a com os respectivos autores.
6. Repete os passos 2 a 5 até que todos os registro retornados pelo passo 1 sejam percorridos pelo algoritmo.

No segundo algoritmo são recuperados os assuntos que são tratados nas publicações, da mesma forma que o anterior são seguidas uma série de passos até a obtenção da informação requerida.

1. Consulta o banco de dados especificamente a tabela *'article'* e retorna os valores do campo *'ee'* que remete para o endereço da publicação.
2. Verifica para qual site o valor do campo *ee* está direcionando.
3. Identifica o site da publicação e redireciona para o algoritmo especifico de cada *website*.
4. Obtém o código HTML da página requisitada e faz a busca dos assuntos da publicação.
5. Após obter as informações da página HTML e relaciona-la com a publicação corrente, persiste a filiação na base de dados associando-a com os respectivos assuntos.
6. Repete os passos 2 a 5 até que todos os registro retornados pelo passo 1 sejam percorridos pelo algoritmo.

No terceiro e último algoritmo de mineração dos campos adicionais o objetivo é recuperar as palavras-chaves (*keywords*) das publicações. A seguir são detalhados os passos seguidos pelo algoritmo:

1. Consulta-se o banco de dados, especificamente a tabela '*article*' e retorna os valores do campo '*ee*' que remete para o endereço da publicação;
2. Verifica-se para qual site o valor do campo *ee* está direcionando;
3. O site da publicação é identificado e redirecionado para o algoritmo específico de cada *website*;
4. Obtém-se o código HTML da página requisitada e faz-se a busca das palavras-chaves da publicação;
5. Após se obter as informações da página HTML e relacioná-las com a publicação corrente, persiste as palavras-chaves na base de dados associando-as com suas respectivas publicações;
6. Repetem-se os passos 2 a 5 até que todos os registro retornados pelo passo 1 sejam percorridos pelo algoritmo.

Na Figura 3.11 são apresentados os campos obtidos via mineração de dados que são os campos à direita, os campos '*keyword*', '*topics*' e '*affiliation*' são obtidos pelos algoritmos de mineração de dados criados especificamente para recuperação dessas informações, já os campos à esquerda são recuperados através da estrutura original do XML provido pelo DBLP.

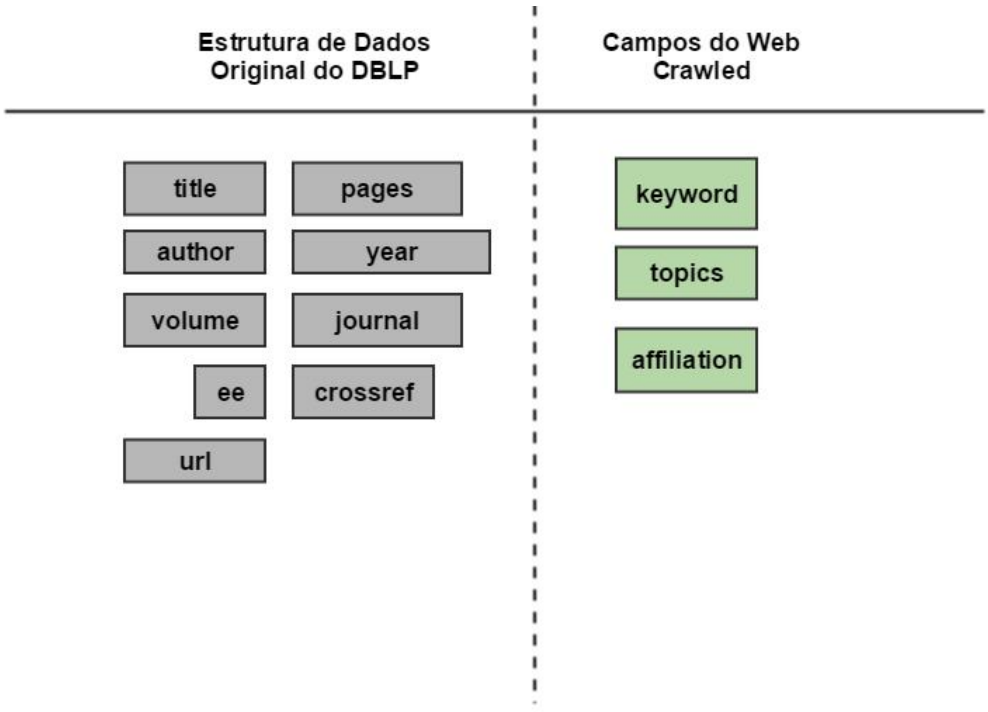

Figura 3.11 - Fontes de dados adicionais do DBLP

3.3 - NORMALIZAÇÃO DA BASE DE DADOS

A etapa de normalização da base de dados consiste na criação de um modelo de entidade e relacionamento genérico que é derivado a partir das estruturas dos modelos do CiteSeerX e do DBLP criados anteriormente.

Nessa etapa as estruturas de dados originais do CiteSeerX e DBLP foram usadas como base para a modelagem de um modelo de entidade e relacionamento normalizado com o propósito de servir de base de dados para a aplicação dos algoritmos deste trabalho.

Desse novo modelo de dados normalizado foram excluídos todos os atributos, entidades e relacionamentos que são inerentes ou exclusivos apenas de uma base de dados específica, no caso deste trabalho as bases do CiteSeerX e do DBLP. E por fim criado um modelo que visa atender as mais diversas fontes de dados bibliométricos.

3.3.1 - MODELO DE ENTIDADE E RELACIONAMENTO OTIMIZADO

Sendo um dos objetivos deste trabalho a construção de um modelo genérico no qual poderão ser usados dados de diversas fontes fez-se necessária a otimização do modelo de entidade e relacionamento criado anteriormente.

O novo modelo de entidade e relacionamento otimizado exclui os campos que não são utilizados nesta pesquisa e também os campos referentes à estrutura do CiteSeerX e do DBLP.

Dessa forma foi criado um novo modelo de entidade e relacionamento otimizado e genérico que pode ser povoado por qualquer fonte de dados acadêmica possibilitando assim o seu reuso em outras pesquisas e aplicações.

A Figura 3.12 apresenta o novo modelo de dados genérico que visa atender a qualquer base de dados acadêmica cumprindo assim um dos objetivos do trabalho que é a criação de um modelo genérico que possa ser aplicado a diversas fontes de informações.

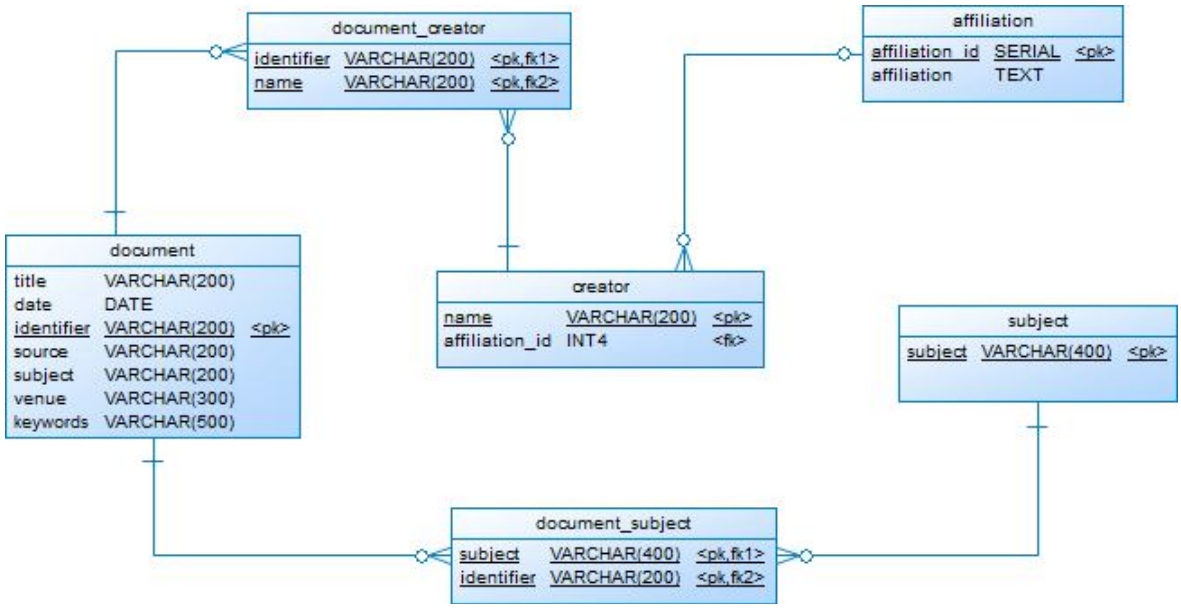

Figura 3.12 - Modelo de Entidade e Relacionamento Otimizado

Com a utilização dos dados contidos no modelo otimizado proposto foram elaborados os modelos para criação dos clusters e redes de comunidades. No entanto, para gerar os clusters e as redes foi necessário preparar a massa de dados a ser usada nas ferramentas de construção dos clusters e das redes de comunidades.

3.4 - CARGA NA BASE MODELO

Após a modelagem e criação da base de dados modelo foi necessário o processo de ETL para carregar essa base com os dados extraídos e minerados do CiteSeerX e DBLP.

Nesse processo de ETL foram necessárias várias etapas de preparação, tratamento e processamento dos dados minerados para possibilitar que os resultados esperados fossem alcançados.

Nessa etapa os dados obtidos pelos processos de mineração e extração foram preparados e carregados na base de dados modelo possibilitando assim que os resultados esperados desta pesquisa fossem alcançados.

É importante salientar que no processo de mineração dos dados foram obtidas diversas informações duplicadas ou até mesmo informações inválidas, que não teriam nenhuma serventia para os fins desta pesquisa.

Para a solução desse problema foi necessário um trabalho para tratamento desses dados, tendo em vista a eliminação de duplicatas e a exclusão de dados corrompidos e/ou até dados inválidos.

A Figura 3.13 apresenta o processo de ETL criado para extrair os dados contidos nas bases do CiteSeerX e DBLP e carrega-los na base de dados modelo.

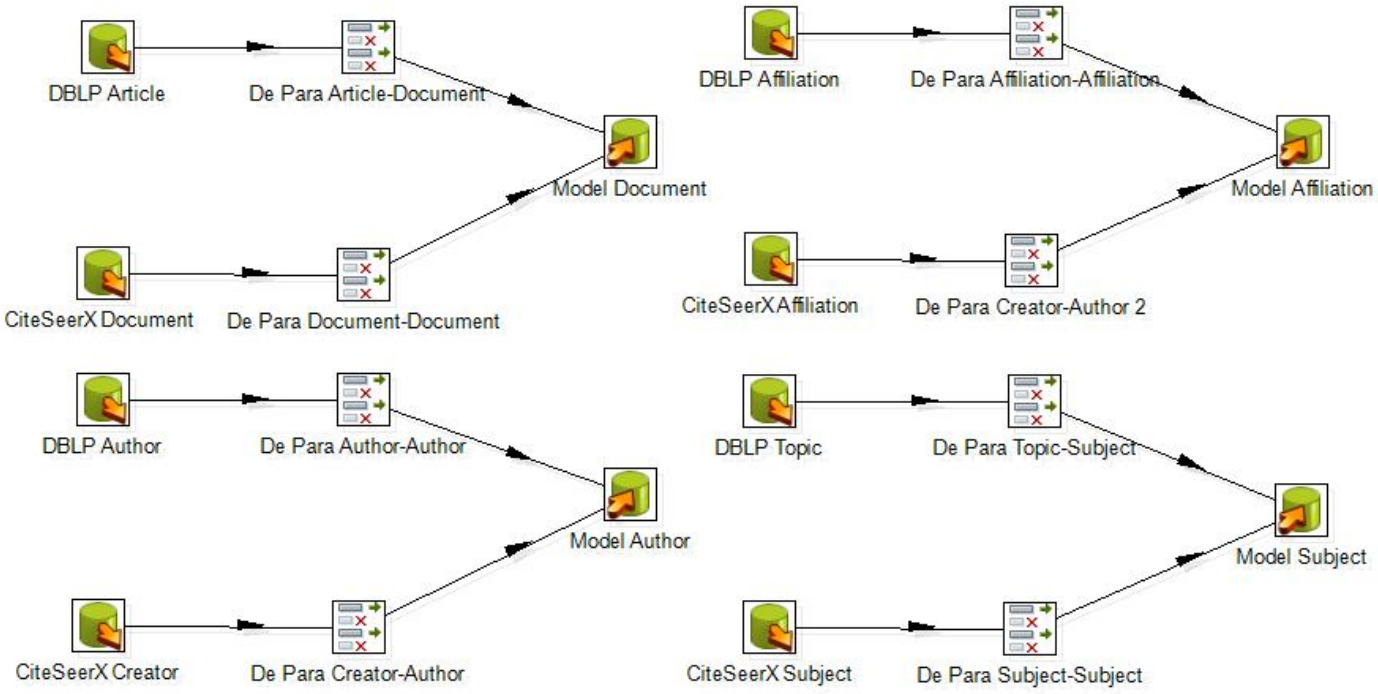

Figura 3.13 - ETL Carga Base Modelo

3.5 - MODELO DE CRIAÇÃO DOS CLUSTERS

Para a criação dos clusters foi usada a ferramenta Weka, desenvolvida na Universidade de Waikato, na Nova Zelândia, e o significado do acrônimo remete para Waikato Environment for Knowledge Analysis (Weka).

Segundo Witten et al (2011) a ferramenta Weka contém um ambiente de trabalho que provê uma coleção de algoritmos no estado da arte de aprendizado de máquina e também ferramentas de pré-processamento de dados. Ela foi projetada para que seja possível testar métodos existentes em várias fontes de dados de formas flexíveis.

O Weka fornece amplo suporte para todo o processo de mineração de dados experimental, incluindo a preparação dos dados de entrada, avaliação da aprendizagem, criação de esquemas estatísticos, visualização dos dados de entrada e do resultado da

aprendizagem. E também fornece uma variedade de algoritmos de aprendizagem de máquina e de clusterização, contendo também uma série de algoritmos distância de medida e possibilitando a combinação de vários algoritmos de clusterização com os algoritmos de distância de medida.

Esse amplo conjunto de recursos é disponibilizado através de uma interface visual comum que possibilita aos usuários uma experiência agradável e com boa usabilidade durante o uso da ferramenta.

Como entrada de dados o Weka recebe um arquivo de formato ARFF (*ARFF Format*). Tal formato consiste em um arquivo de texto que descreve a lista de instâncias que compartilham um conjunto de atributos.

O arquivo ARFF é dividido em duas partes, a primeira contém o cabeçalho que especifica o nome da relação, os atributos da relação, seus determinados tipos e os valores para os tipos com exceção do tipo numérico.

A Figura 3.14 apresenta o cabeçalho do arquivo ARFF, na qual podem ser observadas anotações com @ que identificam os campos. A anotação *@relation* especifica o nome da relação criada e as anotações *@attribute* especificam os atributos contidos na relação e seus possíveis valores.

```
@relation tempo

@attribute perspectiva {ensolarado, nublado, chuvoso}
@attribute temperatura {quente, suave, frio}
@attribute umidade {alta, normal}
@attribute vento {sim, não}
@attribute atividade {sim, não}
```

Figura 3.14 - Cabeçalho do arquivo ARFF

Já o conteúdo é apresentado na Figura 3.15, na qual a anotação *@data* especifica o início da massa de dados. Cada linha na seção dos dados contém um valor para cada atributo separado por vírgula, o conjunto desses valores forma a relação.

No exemplo fornecido a primeira linha diz que o tempo está ensolarado, com a temperatura quente, alta umidade, sem vento e, portanto não aconselhável para prática de atividades.

```
@data
ensolarado,quente,alta,não,não
ensolarado,quente,alta,sim,não
nublado,quente,alta,não,sim
chuvoso,suave,alta,não,sim
chuvoso,frio,normal,não,sim
chuvoso,frio,normal,sim,não
nublado,frio,normal,sim,sim
ensolarado,suave,alta,não,não
ensolarado,frio,normal,não,sim
chuvoso,suave,normal,não,sim
ensolarado,suave,normal,sim,sim
nublado,suave,alta,sim,sim
nublado,quente,normal,não,sim
chuvoso,suave,alta,sim,não
```

Figura 3.15 - Dados do arquivo ARFF

Dada a estrutura do arquivo de entrada da ferramenta mostrou-se necessária a criação de um programa computacional para criar os arquivos com essa estrutura para efetuar o experimento de criação dos clusters.

Na Figura 3.16 é apresentado o pseudocódigo do programa computacional criado para auxiliar na criação dos arquivos ARFF. Nesse pseudocódigo foram usados três atributos (autor, artigo, filiação) para integrarem a relação, no entanto foram criados outros programas baseados nesse pseudocódigo para criar arquivos com os outros atributos usados nos experimentos.

```
/** Classe para o experimento */
class dataset{
    author_id
    article_id
    affiliation_id
}
/** Carrega os dados do cabeçalho*/
[] authors = retrieve_authors_id_from_database();
[] articles = retrieve_articles_id_from_database();
[] affiliations = retrieve_affiliations_id_from_database();

/** Cria o arquivo arff */
arff_file = create_arff_file();

/** Cria a relação */
create_relation(arff_file,'author_article_affiliation')

/** Cria o cabeçalho*/
create_header(authors, arff_file, 'author');
create_header(articles, arff_file, 'article');
create_header(affiliations, arff_file, 'affiliation');

/** Carrega a lista de dados */
retrieve database_data -> dataset

/** Percorre a lista de dados inserindo os registro no arquivo arff  */
for data -> dataset
    create_data(data, arff_file)

/** Salvar o arquivo arff  */
flush(arff_file)
```

Figura 3.16 - Pseudocódigo de criação dos arquivos ARFF

Após a criação dos arquivos de entrada para a ferramenta Weka o passo seguinte foi a criação dos clusters para análise e apresentação dos resultados. No entanto, foi necessário selecionar quais dados seriam usados para gerar os clusters.

Nesse caso, para gerar os clusters foi usada primeiramente toda a massa de dados extraída e minerada do CiteSeerX e do DBLP e posteriormente foram criados clusters apenas com os atributos com maior relevância para esta pesquisa, e filtrados a partir daqueles que possuem o maior número de ocorrências.

Por exemplo, foram criados clusters com os assuntos mais tratados pelas publicações, os autores e as instituições com maior número de publicações. Também foram criadas várias combinações nesse sentido.

3.6 - MODELO DE CRIAÇÃO DOS RANKINGS

Para construir os rankings foram usadas instruções SQL nativas criadas especificamente para esse propósito. Tais instruções foram concebidas usando o modelo de entidade e relacionamento criado nesta pesquisa.

Os rankings foram definidos com o intuito de apresentar uma classificação dos atributos mais relevantes do modelo de dados apresentados. Seguindo a mesma linha de seleção dos atributos dos clusters.

Dessa forma foram criados rankings com as informações referentes aos assuntos, autores, tipos de eventos, filiações e instituições.

Além de criar um ranking específico para cada atributo foram feitas diversas combinações entre os atributos de forma a criar rankings compostos por várias combinações de informações relacionadas, por exemplo, a criação de um ranking composto pelas instituições com mais registros que pesquisam sobre os dez assuntos com maior número de publicações.

3.7 - MODELO DE CRIAÇÃO DAS REDES DE COMUNIDADES

Para construção das redes de comunidades foi usado o software NetDraw, construído por Borgatti (2002) e projetado para criação e visualização de redes sociais de dados.

Borgatti (2002) afirma que enquanto as grandes redes de relacionamento giram em torno de 2 milhões de nós, na prática a maioria dos procedimento do NetDraw são muito lentos para criar redes com mais de 5 mil nós.

No entanto, isso pode variar dependendo de uma análise específica e da baixa densidade da rede. Por exemplo, o grau centralidade pode ser executado em redes de dezenas de milhares de nós, e a maioria das rotinas gráficas tem melhor desempenho quando existem poucos laços, não importando quantos laços existam.

Apesar da limitação na questão do desempenho para geração de redes com milhares de nós essa ferramenta foi escolhida pela facilidade de criação das redes e pela qualidade das redes geradas, além de prover uma boa visualização gráfica.

As redes geradas nesta pesquisa, assim como alguns clusters, giram em torno dos assuntos, instituições e autores com maior número de registros na base de dados criada a partir da mineração dos dados do CiteSeerX.

O tipo de arquivo de entrada para gerar as redes de comunidades é uma matriz cujo exemplo é apresentado na Figura 3.17.

Essa matriz pode ser gerada pela própria ferramenta, no exemplo abaixo as colunas à esquerda definem as instituições, e a primeira linha define os assuntos, nos valores em que a célula está preenchida com zero (0) significa que a instituição não tem publicação no assunto, e onde o preenchimento é dado com o valor um (1) significa que a instituição tem publicação para o assunto em questão.

Criando assim uma matriz binária de informações na qual o 0 (zero) representa o valor falso e o 1(um) representa o valor verdadeiro. Com isso a ferramenta é capaz de construir a rede de relacionamento associando os registros que são comuns.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	ALGORITHMS	DATA MINING	DESIGN	EXPERIMENTATION	INFORMATION	LANGUAGES	MEASUREMENT	NEURAL NETWORKS	PERFORMANCE	RELIABILITY	SECURITY	SIMULATION	XML
2	ENGINEERING DESIGN RESEARCH; LABORATORY;	0	0	1	0	0	0	0	0	0	0	0	1	0
3	SCHOOL OF COMPUTER SCIENCE; UNIVERSITY OF	0	0	0	0	0	0	0	0	1	0	0	0	0
4	SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON	1	1	1	1	0	0	1	1	1	1	1	0	0
5	STRAIGHT-LINE COMPUTATION OF THE - 46 ALLÉE	0	0	0	0	0	0	0	0	0	1	0	0	0
6	CENTRE FOR WIRELESS COMMUNICATIONS;	0	0	0	0	0	0	1	0	0	0	0	1	0
7	UNIVERSIDADE DOS AÇORES - APARTADO 1422,	0	0	0	0	0	0	0	0	0	0	0	0	0
8	ZURICH, SCHWEIZ - ECOLE SUISSE POLYTECHNIQUE	0	0	0	0	0	0	0	0	0	0	0	0	0
9	DREYER URGENT CARE CENTER (DREYER UCC) IS	0	0	0	0	0	0	0	0	0	0	0	0	0
10	DEFENSE MODELING AND SIMULATION OFFICE - 1901 N.	0	0	0	0	0	0	0	0	0	0	0	0	0
11	FOOD CONSUMPTION AND NUTRITION DIVISION;	0	0	0	0	0	0	0	0	0	0	0	0	0
12	ECOLE NORMALE SUPÉRIEURE DE LYON - 46 ALLÉE	0	0	0	0	0	0	0	0	0	0	0	1	0
13	PROCEEDINGS FORMATTING TEAM - JOHN W.	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1 IICM- SOFTWARETECHNOLOGY, GRAZ UNIVERSITY	1	0	0	0	0	0	0	0	0	0	0	0	0
15	MODEL VERIFICATION AND VALIDATION - JOHN S.	0	0	0	0	0	0	0	0	0	0	0	0	0
16	FACULTY OF ELECTRICAL ENG. & INFORMATION	0	0	0	0	0	0	0	0	0	0	0	0	0
17	TEL-AVIV UNIVERSITY, TEL-AVIV, ISRAEL, AND	1	0	0	0	0	0	0	0	0	0	0	0	0
18	CWI - P.O. BOX 94079, 1090 GB AMSTERDAM, THE	0	0	0	0	0	1	0	0	0	0	0	0	0
19	1; DEPARTMENT OF COMPUTER SCIENCE; 2; SCHOOL	0	0	1	0	0	0	0	0	1	0	0	0	0
20	ENVIRONMENT AND PRODUCTION TECHNOLOGY	0	0	0	0	0	0	0	0	0	0	0	0	0
21	MOTOR VEHICLES; SIMULATION AS A PRIMARY TOOL	0	0	0	0	0	0	0	0	0	0	0	0	0
22	DARMSTADT UNIVERSITY OF TECHNOLOGY, DEPT. OF	0	0	0	0	0	1	0	0	0	0	1	0	0
23	; NOTES; © 2005 COLD SPRING HARBOR LABORATORY	1	0	1	0	0	0	0	0	0	0	0	0	0
24	I DEPARTMENT OF ELECTRICAL ENGINEERING &	0	0	0	1	0	0	1	0	0	0	0	0	0
25	DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY	1	0	0	0	0	0	0	0	1	0	1	0	0
26	COMPUTER ENGINEERING LABORATORY;; ELECTRICAL	0	0	0	0	0	0	0	0	1	0	0	0	0

Figura 3.17 - Matriz de filiação e assunto

Dessa forma a ferramenta cria as redes de comunidades que irão ser apresentadas, analisadas e discutidas em detalhes no próximo capítulo.

3.8 - SÍNTESE DO CAPÍTULO

Esse capítulo teve como objetivo apresentar os métodos e técnicas utilizados para obtenção dos modelos e dados produzidos por este trabalho. Nele foram ilustradas as técnicas utilizadas para a mineração dos dados requeridos, os modelos de entidade e relacionamento criados, e os algoritmos criados para mineração dos dados adicionais que não foram obtidos através da primeira mineração.

Na segunda parte foram apresentados os modelos para geração dos resultados esperados. Nesses modelos estão contidos o modelo de entidade e relacionamento otimizado, que visa ser genérico e aplicável a qualquer base de dados acadêmica, e os modelos para criação dos clusters que foram criados a partir da ferramenta Weka, rankings que foram criados com o auxílio das consultas SQL concebidas exclusivamente para este propósito, e as redes de relacionamento que foram geradas com o modelo proposto pela ferramenta NetDraw.

4. RESULTADOS OBTIDOS

Com a aplicação do modelo de extração transformação e carga auxiliado do processo de mineração de dados foi possível obter uma vasta massa de dados para a aplicação dos algoritmos de clusters, criação dos rankings e identificação das redes de relacionamento.

A Tabela 4.1 apresenta a quantidade de registros obtidos do CiteSeerX para cada atributo que será usado na criação dos objetos de análise deste trabalho.

Tabela 4.1 - Quantidade de registros do CiteSeerX

Quantidade de Registros do CiteSeerx	
Autores	206.666
Publicações	112.629
Assuntos	72.624
Instituições	66.324

Na Tabela 4.2 são apresentados os números de registros obtidos do DBLP para cada atributo que servirá de insumo para a criação dos resultados.

Tabela 4.2 - Quantidade de registros do DBLP

Quantidade de Registros do DBLP	
Autores	116.150
Publicações	1.391.558
Assuntos	4.614
Instituições	892
Palavra-Chave	518

A partir dos registros obtidos foram criados clusters com auxílio da ferramenta Weka a título de ilustração e comprovação do modelo proposto. Sendo que para a geração dos clusters e redes de relacionamento foram recuperadas informações da base de dados modelo por amostragem em que os atributos com maior número de registros foram selecionados para a formação dos clusters.

É importante destacar que os resultados apresentados nos clusters, rankings e redes de relacionamento não refletem a realidade do cenário acadêmico científico mundial, e devem ser analisados apenas em forma (estrutura) e não em conteúdo.

Portanto foram criados dois clusters a partir dos dados do CiteSeerX e um cluster para os dados do DBLP, sendo que os clusters gerados a partir dos dados do CiteSeerX possuem os atributos de filiação, tipo de evento, autor e assunto, e o gerado a partir do DBLP contém os atributos de filiação, autor e assunto.

4.1 - CLUSTERS

O primeiro cluster foi criado a partir do algoritmo *SimpleKmeans* (implementação da ferramenta Weka do algoritmo *Kmeans*) configurado para criar quatro clusters usando a distância euclidiana.

Nesse cluster os atributos usados são tipo de evento, filiação e autor. Onde tipo de evento remete as conferências e revistas nas quais os trabalhos são publicados, filiação que são as instituições que os autores são vinculados, e os próprios autores dos trabalhos.

A Tabela 4.3 apresenta o centroide do cluster criado, ela mostra o centroide de cada cluster para cada atributo e o centroide principal da massa de dados utilizada.

Tabela 4.3 - Primeiro cluster CiteSeerX - tipo de evento, filiação e autor

Cluster centroids	**Cluster#**				
Full Data	**0**	**1**	**2**	**3**
299 (100%)	128 (43%)	119 (40%)	5 (2%)	47 (16%)
Attribute					
venue_type	CONFERENCE	JOURNAL	CONFERENCE	CONFERENCE	CONFERENCE
affiliation	School of Computer Science; Carnegie Mellon University - Pittsburgh, PA	Department of Electrical; Computer Engineering; University of Waterloo	School of Computer Science; Carnegie Mellon University - Pittsburgh	School of Computer Science; Carnegie Mellon University - Pittsburgh	Engineering Design Research; California Institute of Technology
author	Student Member	Student Member	Michael K. Reiter	Orna Grumberg	Ph. D

A Figura 4.1 apresenta a visualização gráfica do cluster gerado dando ênfase à filiação e ao tipo de evento. Nela é possível observar que diferentes universidades,

Carnigie Mellon University e *University of Waterloo*, possuem a maior quantidade de registros para os tipos de eventos, conferências e revistas respectivamente, o que mostra que apesar da universidade *Carnigie Mellon* possuir a maior quantidade de registros a sua representatividade de publicações em revistas é consideravelmente inferior ao número de publicações em conferências.

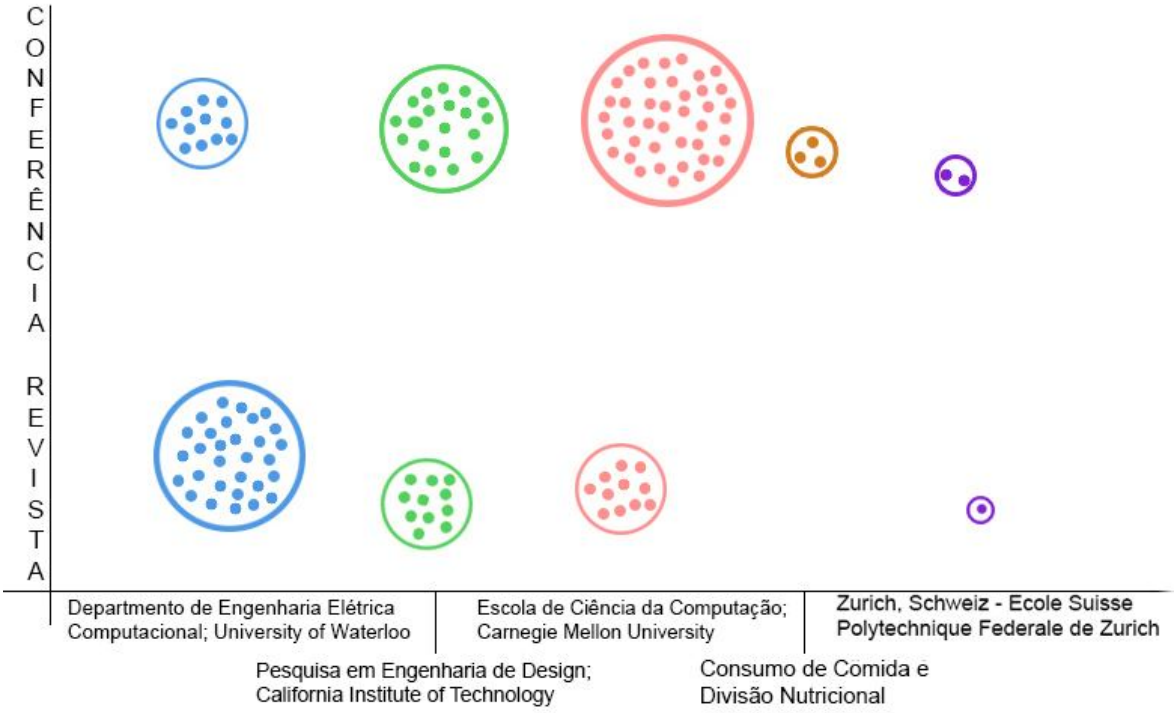

Figura 4.1 - Cluster CiteSeerX - Filiação X Tipo de Evento

Na Figura 4.2 é possível visualizar a representação do cluster com a perspectiva de número de autores por instituições. Nessa visualização as publicações em conferências e revistas são somadas ao objetivo de apresentar o montante de autores que compõe trabalhos científicos pelas respectivas instituições.

Como resultado a Universidade *Carnegie Mellon* possui a maior quantidade de autores seguida pela Universidade de *Waterloo*, o que condiz com a divisão dos centroides dos clusters gerados.

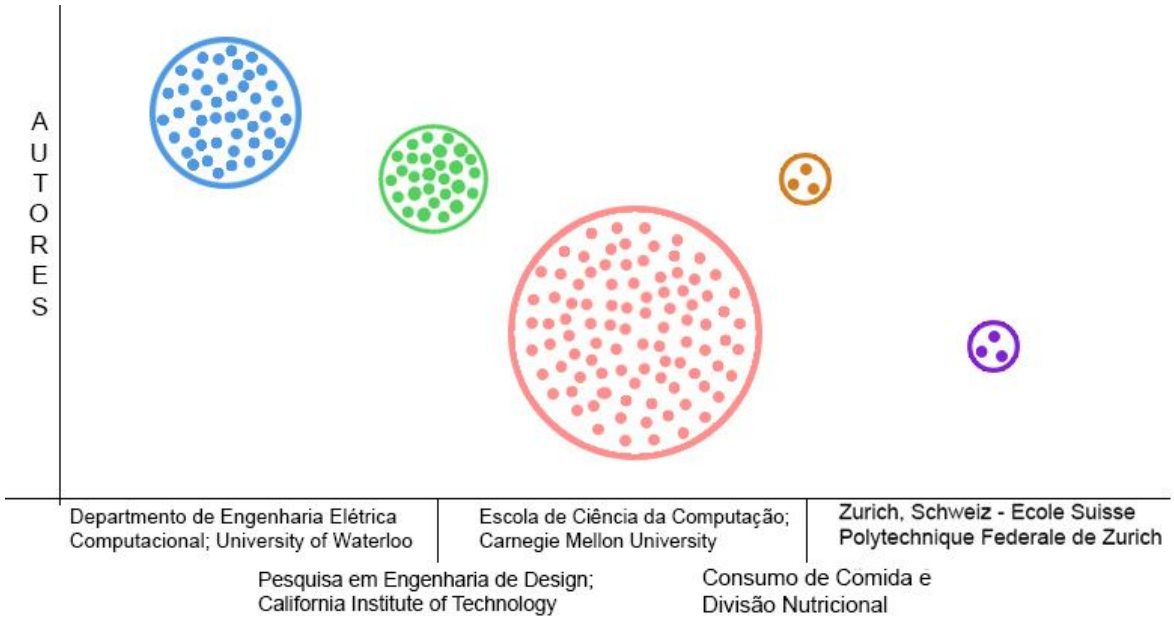

Figura 4.2 - Cluster CiteSeerX - Filiação X Autor

Na Figura 4.3 é apresentada a visualização do cluster com a proporção de autores para conferências e revistas.

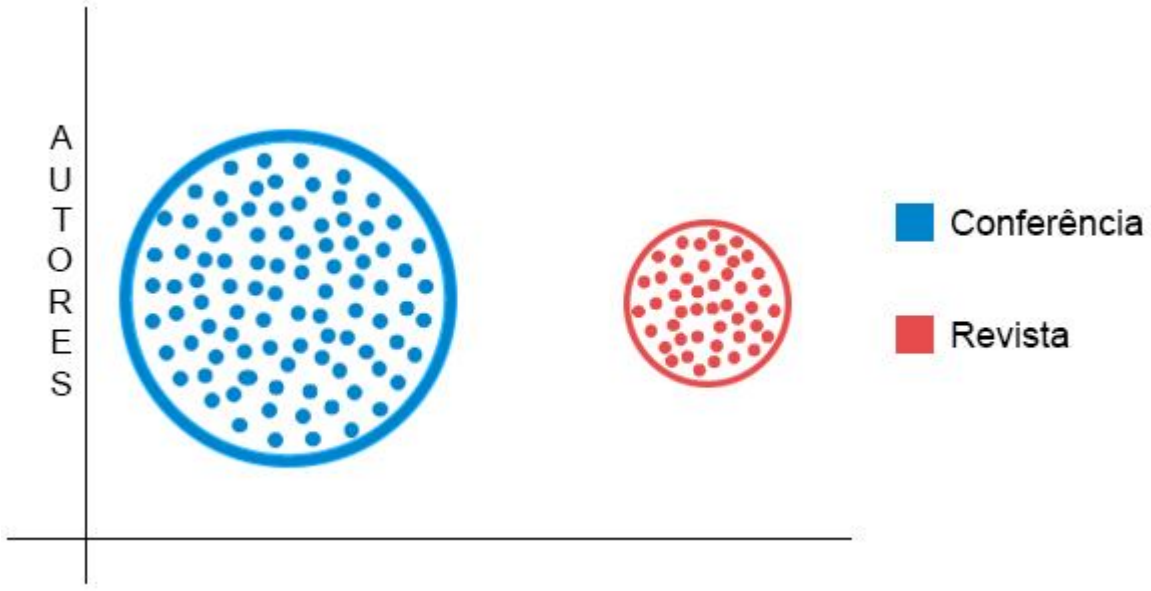

Figura 4.3 - Cluster CiteSeerX - Autor X Tipo de Evento

A Tabela 4.4 apresenta o centroide do segundo cluster que foi gerado a partir dos quarenta assuntos com maior número de registros com os filtros de conferências e revistas para o tipo de evento e as dez instituições com mais ocorrências e seus respectivos autores.

Nesse cluster, assim como no anterior, foi usado o algoritmo *kmeans*, com a diferença na configuração do número de clusters, sendo que neste foram usados dois clusters ao invés de quatro, a distância de medida adotada continuou sendo a euclidiana.

Tabela 4.4 - Cluster CiteSeerX tipo de evento, filiação, autor e assunto

Cluster centroids:	**Cluster#**	
Full Data	**0**	**1**
177 (100%)	57 (32%)	120 (68%)
Attribute		
venue type	CONFERENCE	JOURNAL	CONFERENCE
affiliation	SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY	DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BOLOGNA	SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY
author	SCOTT SHENKER	GEORGE KARYPIS	SCOTT SHENKER
subject	ALGORITHMS	ALGORITHMS	DESIGN

A Figura 4.4 apresenta o cluster com o comparativo dos assuntos publicados em conferências e revistas. Nessa visualização existe certa simetria entre os registros, de forma que as conferências possuem maior quantidade de registros para cada assunto e as revistas mantêm a proporção em uma escala menor.

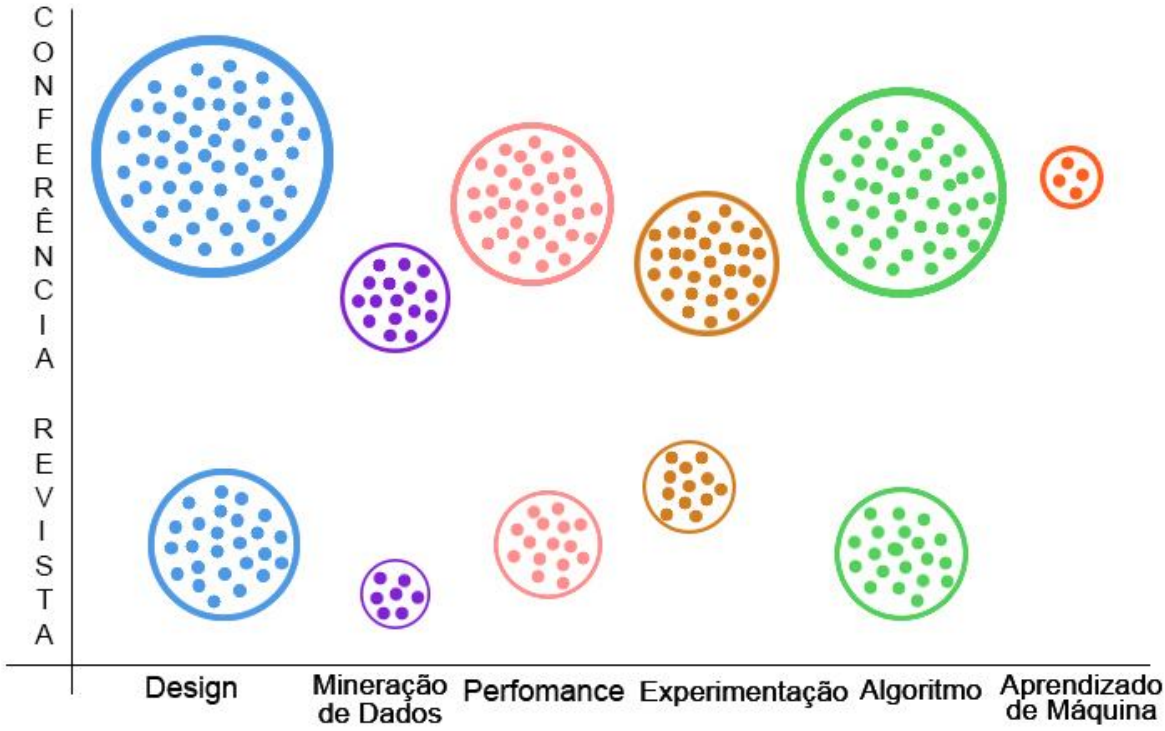

Figura 4.4 - Cluster CiteSeerX - Tipo Evento X Assunto

A Figura 4.5 apresenta aglomerado de autores que publicam nos respectivos assuntos. Nessa visão os artigos publicados em conferências e revistas são somados para se chegar ao acumulado de autores para cada assunto, em que são apresentados os seis assuntos mais recorrentes na massa de dados.

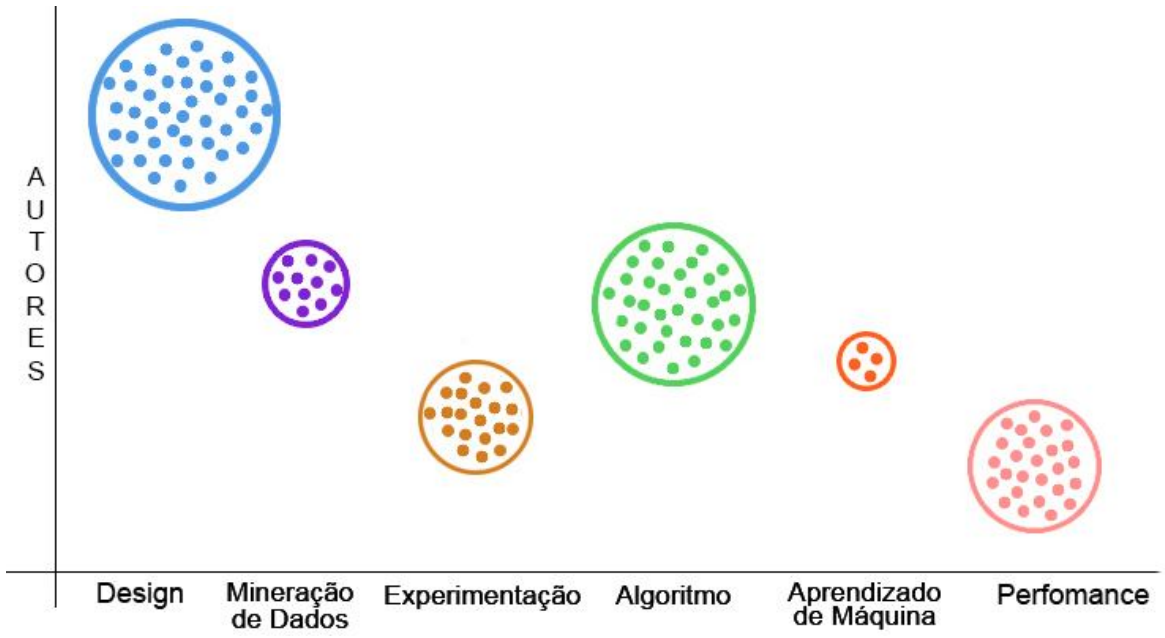

Figura 4.5 - Cluster CiteSeerX - Autor X Assunto

O cluster da Figura 4.6 mostra as instituições que mais publicam em determinados assuntos, sem a distinção entre os tipos de evento. Onde a instituição com maior número de registros no cluster, a Universidade *Carnigie Mellon* tem o maior número de publicações em 3 dos 5 assuntos contidos nessa visão.

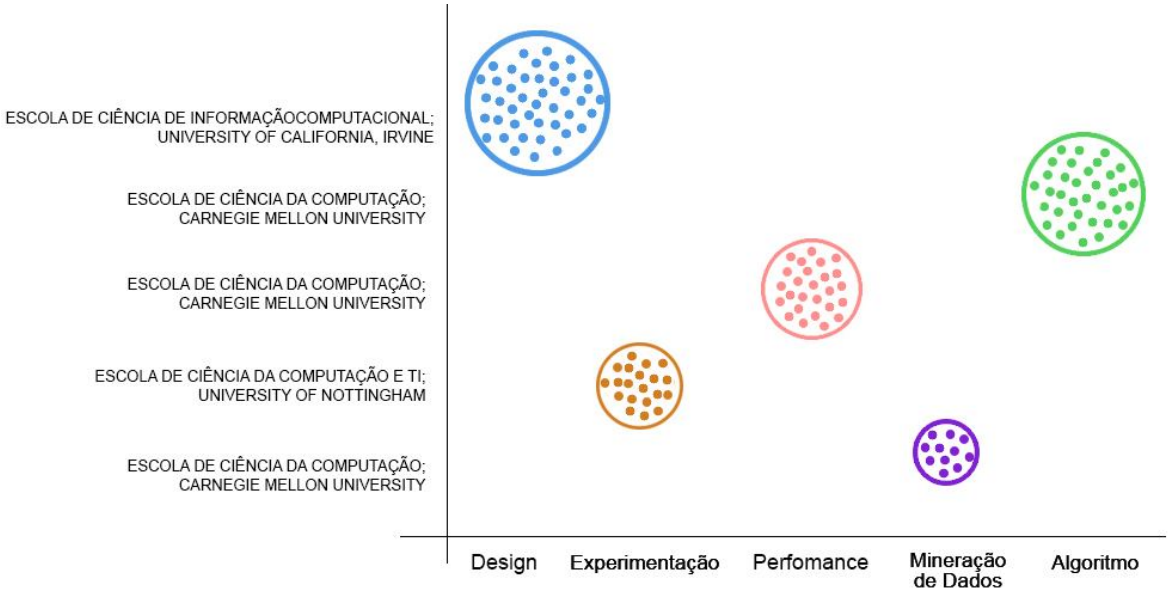

Figura 4.6 - Cluster CiteSeerX - Filiação X Assunto

A Tabela 4.5 apresenta o centroide para o cluster gerado a partir dos dados do DBLP com os atributos de autor, filiação e assunto. Na sua configuração foram usados seis (6) clusters e a distância de medida euclidiana.

Tabela 4.5 - Clusters DBLP - Filiação, assunto e autor

Cluster centroids	**Cluster#**					
Full Data	**0**	**1**	**2**	**3**	**4**	**5**
4027 (100%)	3540 (88%)	175 (4%)	37 (1%)	34 (1%)	43(1%)	198 (5%)
Attribute						
affiliation	German Aerosp. Center, Remote Sensing Technol. Inst., Wessling, Germany	German Aerosp. Center, Remote Sensing Technol. Inst., Wessling, Germany	Dept. of Electr. Eng., Tsinghua Univ., Beijing, China	Cyber Security & Intell. Dept., A*STAR Inst. for Infocomm Res., Singapore, Singapore	Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK	Coll. of Comput. Sci., Zhejiang Provincial Key Lab. of Service Robot., Zhejiang Univ., Hangzhou, China	Inst. of Comput., Univ. of Campinas, Campinas, Brazil
subject	Data Structures, Cryptology and Information Theory	Data Structures, Cryptology and Information Theory	Mathematics of Computing	Delays	Algorithm Analysis and Problem Complexity	Vectors	Data Structures, Cryptology and Information Theory
author	Leah Epstein	Leah Epstein	K. Subramani	Dinil Mon Divakaran	Gregory Gutin	Chun Chen	Siome Goldenstein

A Figura 4.7 apresenta o cluster com a filiação no eixo Y e o autor no eixo X. O resultado mostra a filiação "*German Aerosp. Center, Remote Sensing Technol. Inst., Wessling, Germany*" e o autor *Leah Epstein* com o maior número de registros no cluster.

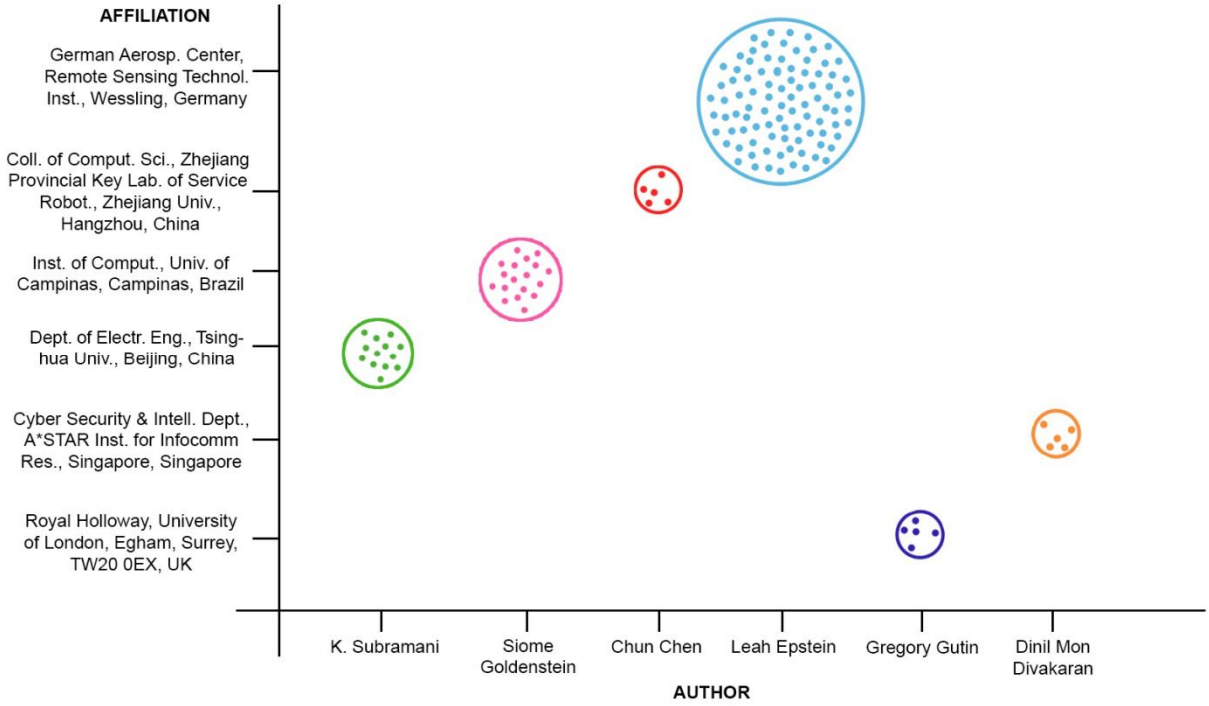

Figura 4.7 - Cluster DBLP - Filiação X Autor

A Figura 4.8 apresenta o cluster com os assuntos no eixo Y e os autores no eixo X. Nessa visão os assuntos com maior número de registros são *'Data Structures, Cryptology and Information Theory', 'Vectors', 'Algorithm Analysis and Problem Complexity', 'Delays and Mathematics of Computing'*.

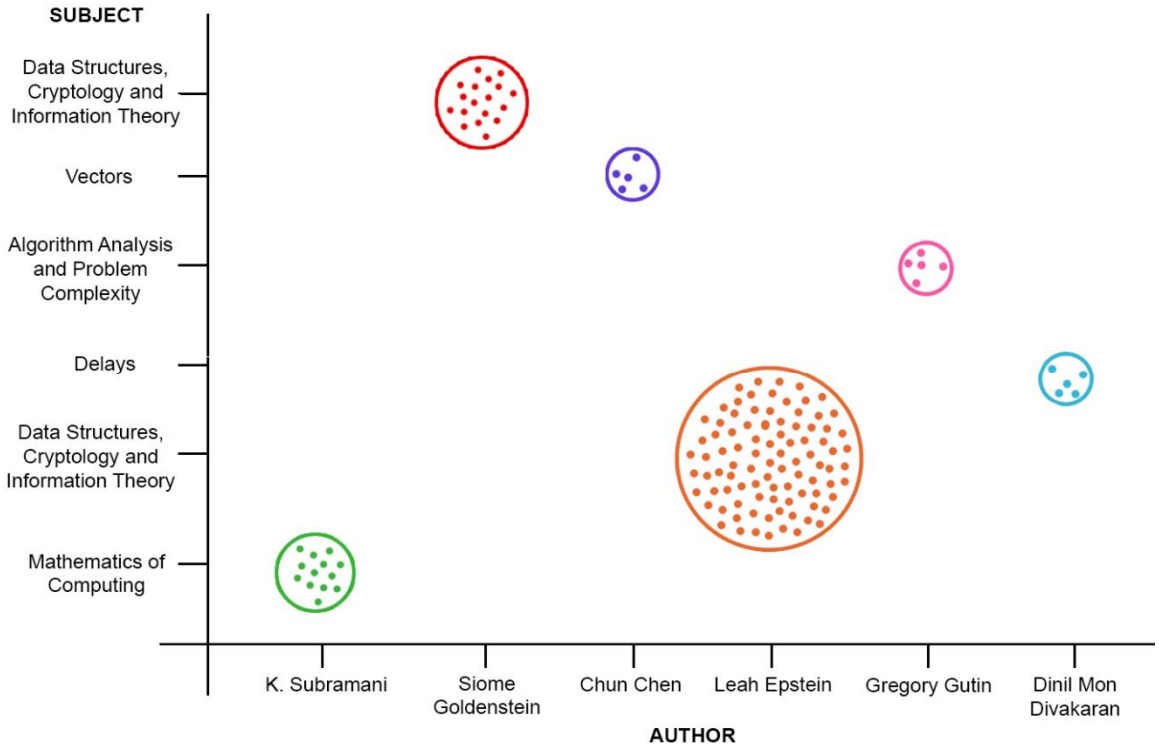

Figura 4.8 - Cluster DBLP - Assunto X Autor

A Figura 4.9 apresenta o cluster com os assuntos no eixo X e as filiações no eixo Y. Essa última apresentação do cluster gerado a partir dos dados do DBLP se mostra bastante similar à anterior, se diferenciando pelo foco na filiação ao invés de focar o autor.

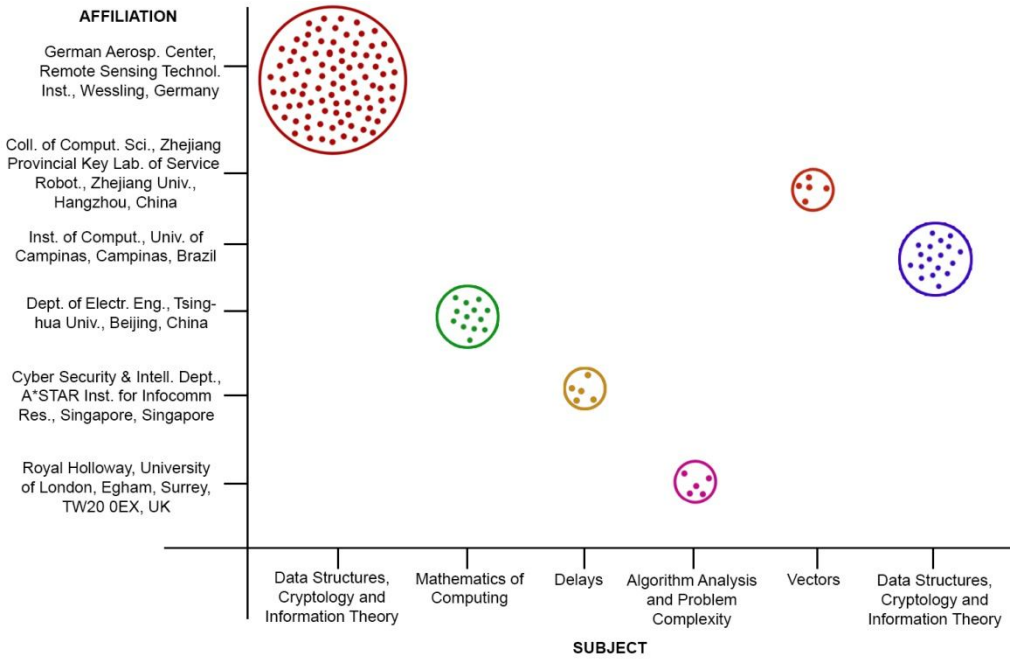

Figura 4.9 - Cluster DBLP - Filiação X Assunto

4.2 - RANKINGS

Para complementar a análise sobre os clusters foi gerada uma série de rankings que mostram, por exemplo, qual instituição tem o maior número de registros dentro dos dados obtidos para a pesquisa, os assuntos mais abordados e publicados em conferências e revistas.

A Tabela 4.6 apresenta o ranking com o número de ocorrências para os tipos de eventos encontrados na base de dados do CiteSeerX.

Tabela 4.6 - Ranking CiteSeerX de tipo de evento

Tipo de Evento	Número de registros
CONFERENCE	25302
JOURNAL	8510
TECHREPORT	898
PROCEEDINGS	11
WORKSHOP	5
BOOK	3
SUMMER SCHOOL	
POSTER	2
INPROCEEDINGS	2
PH,D, THESIS	2
M.SC. THESIS	1
GI WORKSHOP	1
ELECTRONIC NOTES	
IN	1
WORKSHOP	
PROCEEDINGS	1
LECTURE NOTES IN	
COM	1
BOOK CHAPTER	1
PH.D. DISSERTATION	1
SIRS 98	1
ARTICLE	1
THESIS	1
INTERNATION	
CONFEREN	1
RESEARCH REPORT	1
MASTER'S THESIS	1
TECHNICAL REPORT	1
REPORT	1

Na Tabela 4.7 são apresentadas as instituições com maior número de registros no total, não aplicando nenhum filtro.

Tabela 4.7 - Ranking CiteSeerX dos Institutos

Institute	**Count**
ENGINEERING DESIGN RESEARCH; LABORATORY; CALIFORNIA INSTITUTE OF TECHNOLOGY - PASADENA, CALIFORNIA 91125	460
SCHOOL OF COMPUTER SCIENCE; UNIVERSITY OF BIRMINGHAM - EDGBASTON, B15 2TT, UK	364
SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 15213	269
STRAIGHT-LINE COMPUTATION OF THE - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE	269
CENTRE FOR WIRELESS COMMUNICATIONS; DEPARTMENT OF ELECTRICAL; COMPUTER ENGINEERING; UNIVERSITY OF WATERLOO	260
UNIVERSIDADE DOS AÇORES - APARTADO 1422, 9501-801 PONTA DELGADA, PORTUGAL	234
ZURICH, SCHWEIZ - ECOLE SUISSE POLYTECHNIQUE FEDERALE DE ZURICH	232
DREYER URGENT CARE CENTER (DREYER UCC) IS LOCATED IN - AURORA, ILLINOIS.	225
DEFENSE MODELING AND SIMULATION OFFICE - 1901 N. BEAUREGARD STREET; ALEXANDRIA, VA 22311	143
FOOD CONSUMPTION AND NUTRITION DIVISION; INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE	131
ECOLE NORMALE SUPÉRIEURE DE LYON - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE	123
PROCEEDINGS FORMATTING TEAM - JOHN W. LOCKHART, RED HAT, INC.	117

1 IICM- SOFTWARETECHNOLOGY, GRAZ UNIVERSITY OF TECHNOLOGY, - INFFELDGASSE 16B/I A-8010 GRAZ, AUSTRIA	115
MODEL VERIFICATION AND VALIDATION - JOHN S. CARSON, II	113
FACULTY OF ELECTRICAL ENG.; INFORMATION TECHNOLOGY; SLOVAK UNIVERSITY OF TECHNOLOGY	112

A Tabela 4.8 apresenta os assuntos com mais ocorrências sem a distinção do tipo de evento em que o trabalho foi publicado.

Tabela 4.8 - Ranking CiteSeerX dos assuntos

Subject	**Count**
DESIGN	846
ALGORITHMS	670
PERFORMANCE	584
CONTENTS	577
EXPERIMENTATION	432
SECURITY	342
GENERAL TERMS	333
CATEGORIES AND SUBJECT DESCRIPTORS	327
MEASUREMENT	291
SIMULATION	258
DATA MINING	236
LANGUAGES	204
HUMAN FACTORS	196
CLASSIFICATION	188
INFORMATION RETRIEVAL	186

Na Tabela 4.9 são agrupados os atributos tipo de evento e assunto e ranqueados conforme a quantidade de registros de cada agrupamento.

Tabela 4.9 - Ranking CiteSeerX de tipo de evento por assunto

Venue type	**Subject**	**Count**
CONFERENCE	DESIGN	248
CONFERENCE	PERFORMANCE	172
CONFERENCE	ALGORITHMS	169
CONFERENCE	EXPERIMENTATION	130

CONFERENCE	SECURITY	99
JOURNAL	ALGORITHMS	93
JOURNAL	CATEGORIES AND SUBJECT DESCRIPTORS	92
CONFERENCE	MEASUREMENT	82
CONFERENCE	GENERAL TERMS	80
CONFERENCE	CATEGORIES AND SUBJECT DESCRIPTORS	71
JOURNAL	DESIGN	70
JOURNAL	GENERAL TERMS	56
JOURNAL	PERFORMANCE	37
CONFERENCE	SIMULATION	37
JOURNAL	EXPERIMENTATION	33
CONFERENCE	CONTENTS	32
JOURNAL	CONTENTS	28
JOURNAL	SECURITY	23
JOURNAL	MEASUREMENT	20
JOURNAL	SIMULATION	11

A Tabela 4.10 apresenta o ranking das trinta filiações com maior recorrência nos dados extraídos do DBLP.

Tabela 4.10 - Rankin DBLP de filiação

Affiliation	**Count**
German Aerosp. Center, Remote Sensing Technol. Inst., Wessling, Germany	81
Dept. of Electr. Eng., Tsinghua Univ., Beijing, China	22
Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark	17
Dept. of Comput. Sci. & Eng., Shanghai Jiao Tong Univ., Shanghai, China	15
Dept. of Electron. Eng., City Univ. of Hong Kong, Hong Kong, China	12
Department of Mathematics, University of Haifa, 31905, Haifa, Israel	12
Inst. of Comput., Univ. of Campinas, Campinas, Brazil	11
School of Electrical Engineering, Tel Aviv University, 69978, Tel Aviv, Israel	10
Sch. of Electr. Eng., Southwest Jiaotong Univ., Chengdu, China	10
Yahoo! Research, Haifa, Israel	10
Ecole Polytech. Fed. de Lausanne, Lausanne, Switzerland	10

Coll. of Electr. & Inf. Eng., Hunan Univ., Changsha, China	10
Division of Information System Design, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama, 350-0394, Japan	9
Dept. of Eng., Roma Tre Univ., Rome, Italy	9
Dept. of Comput. Sci., Guangxi Normal Univ., Guilin, China	9
Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR	9
State Key Lab. of Integrated Services Networks, Xidian Univ., Xi'an, China	9
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada	9
School of Computer Science, University of Waterloo, 200 University Ave. W, Waterloo, ON, N2L3G1, Canada	9
Department of Statistics, The George Washington University, Washington, DC, 20052, USA	8

A Tabela 4.11 apresenta o ranking com os vinte autores com maior número de registros extraídos do DBLP.

Tabela 4.11 - Ranking DBLP de autores

Author	**Count**
H. Vincent Poor	812
Chin-Chen Chang	545
Witold Pedrycz	535
Yan Zhang	448
Lajos Hanzo	438
Wei Wang	434
Xiaodong Wang	423
Mohamed-Slim Alouini	416
Noga Alon	408
Azriel Rosenfeld	383
Guanrong Chen	372
Georgios B. Giannakis	371
Wei Liu	369
Ronald R. Yager	366
Yong Wang	366
Loet Leydesdorff	357
Jing Li	343

Lei Wang	340
Wei Zhang	340
Jian Li	339

A Tabela 4.12 apresenta o ranking com os assuntos que são mais tratados na base de dados extraída do DBLP.

Tabela 4.12 - Ranking DBLP de assuntos

Subject	**Count**
Data Structures, Cryptology and Information Theory	125
Mathematics of Computing	117
Computer Systems Organization and Communication Networks	117
Algorithm Analysis and Problem Complexity	117
Algorithms	117
Theory of Computation	117
Atmospheric modeling	94
Synthetic aperture radar	81
Tomography	81
Geometry	81
Orbits	81
Azimuth	81
Three-dimensional displays	81
Optimization	54
Robustness	51
Load modeling	50
Microgrids	47
Power generation	44
Algorithm design and analysis	42
Security	40

A Tabela 4.13 apresenta as vinte palavras-chaves mais recorrentes na base de dados minerada do DBLP.

Tabela 4.13 - Ranking DBLP de palavras-chave

Keyword	**Count**
TerraSAR-X	81
stereo SAR	81
geodetic SAR tomography	81
geodetical fusion	81
SAR tomography	81
Absolute positioning	81
SAR geodesy	81
synthetic aperture radar (SAR)	81

smart grid	23
microgrid	17
uncertainty	17
reliability	17
energy efficiency	15
Biometrics	15
demand response (DR)	15
energy harvesting	14
Compiler	13
virtual power plant (VPP)	12
multimedia phylogeny	11
Image forensics	11

4.3 - COMUNIDADES

A Figura 4.10 apresenta o gráfico de rede de relacionamento com as dez instituições com maior número de registros e os quinze assuntos melhor ranqueados. Como pode ser observado existe apenas um assunto sem relacionamento com as instituições, todos os outros assuntos contêm relacionamento. As dez instituições apresentadas no gráfico são diferentes das instituições apresentadas no ranking, isso acontece porque grande quantidade de registros obtidos não contém informações sobre os assuntos inerentes a cada publicação.

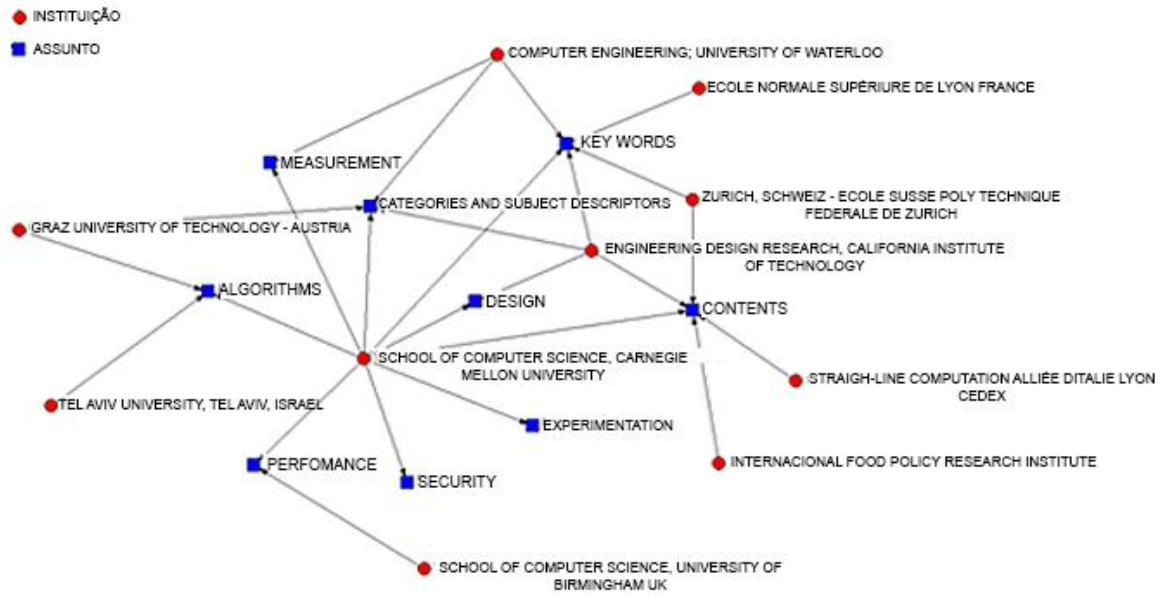

Figura 4.10 - Rede de comunidade de assuntos

Na Figura 4.11 é apresentada a rede de relacionamento que aborda a colaboração existente entre instituições. Para criar essa rede de colaboração foram selecionados os institutos que se relacionam com os autores com maior número de registros na massa de dados.

Nota-se que mais de 80% dos institutos são provenientes dos Estados Unidos e os restantes são de diferentes regiões por todo o mundo, como, Alemanha, Israel, Polônia, Índia e Brasil.

Os resultados mostram que há significativa rede de colaboração entre a Universidade de Princeton, nos Estados Unidos, e a Universidade de Tel Aviv, em Israel, e no nó que contém essas duas universidades são encontradas colaborações entre institutos de diferentes países e continentes.

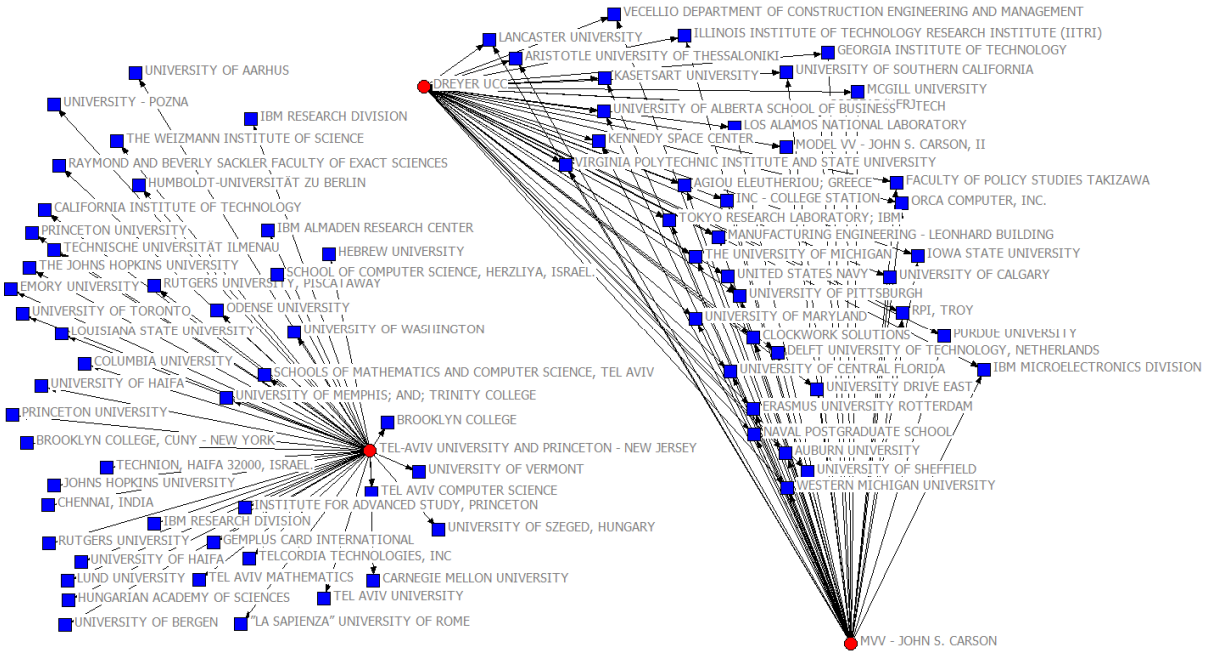

Figura 4.11 - Rede de comunidade de afiliações

Na Figura 4.12 foi criada a rede de colaboração abordando os quatro autores com maior número de registro para formar a base dos nós, e os autores que trabalham em colaboração com os melhores ranqueados.

Foi descoberto que três dos quatro autores com maior número de registros compartilham praticamente a mesma rede de colaboração, e um deles tem uma rede de colaboração totalmente separada dos outros três.

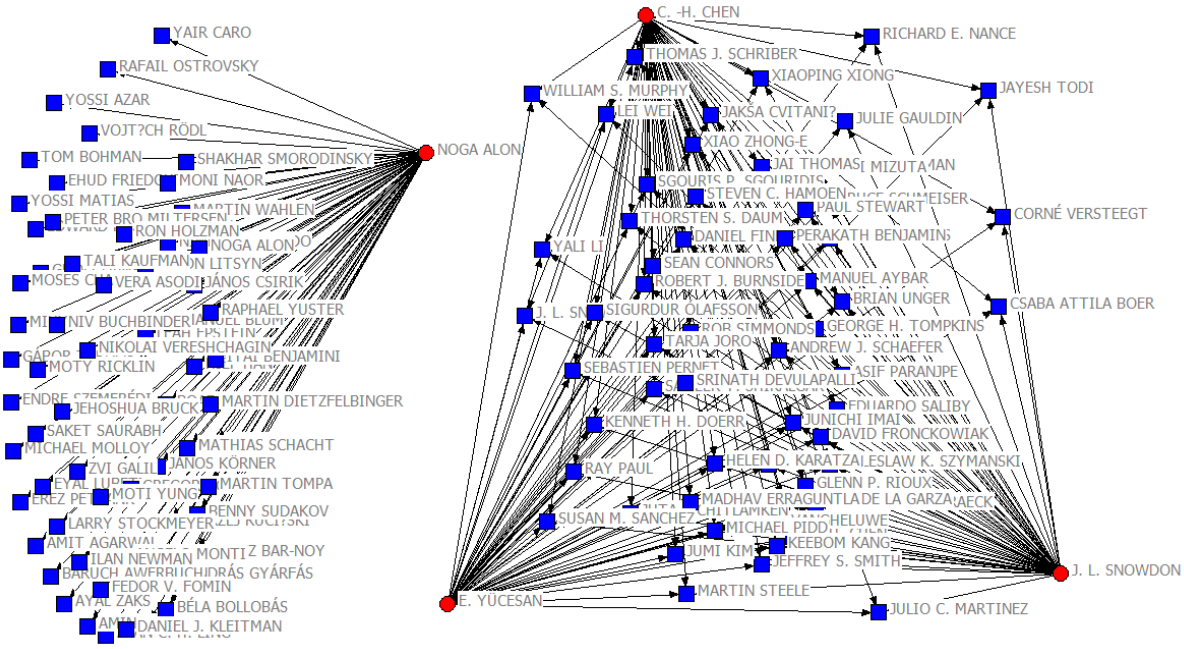

Figura 4.12 - Rede de comunidade de autores

4.4 - SÍNTESE DO CAPÍTULO

Esse capítulo teve como objetivo a exploração dos resultados obtidos através da aplicação do modelo proposto no capítulo anterior.

Foram apresentados os resultados nos quais foram identificadas as instituições, autores e assuntos que são mais recorrentes na comunidade acadêmica dentro do escopo dos dados obtidos. Foram criados rankings com diversas combinações de atributos. E por fim foram identificadas as redes de relacionamento das instituições, autores e assuntos com maior número de registros na base de dados minerada.

5 - CONCLUSÕES

O trabalho teve como objetivo a criação de um modelo genérico de análise de dados que pudesse explorar diversas bases de dados acadêmicas em busca de conhecimento sobre os assuntos mais pesquisados, as comunidades de autores e suas respectivas instituições. Para isso, foi tomado como base o modelo de dados criado a partir da estrutura do CiteSeerX e do DBLP, e com isso criado um modelo abrangente que exclui as particularidades dos modelos em questão com o objetivo de criar um modelo que seja compatível com diversas entradas de dados.

O modelo criado abrange as etapas de extração dos dados das bases bibliométricas, mineração dos dados complementares que não são contemplados nas estruturas originais disponibilizadas pelo CiteSeerX e DBLP, criação da base de dados modelo, carga na base de dados modelo, e por fim a geração dos resultados.

Na etapa de obtenção dos dados para a execução do experimento proposto de criação dos clusters, rankings e redes de relacionamento foi usada a API, provida pelo CiteSeerX, e os arquivos XML disponibilizados para download pelo DBLP. E para os dados complementares foram criados algoritmos (*webcrawlers*) para mineração dessas informações.

Com a etapa de aquisição dos dados concluída foi possível criar, a partir das estruturas de dados fornecidas, um modelo de dados abrangente a várias entradas de dados provindas de diferentes bases de dados acadêmicas. E a partir desse modelo foi criado o processo de carga e posteriormente a geração dos resultados.

O objetivo principal da apresentação dos resultados gerados a partir dos dados minerados nesta pesquisa é exemplificar as conclusões e análises que podem ser geradas a partir da aplicação do modelo. Dessa forma, é importante se atentar mais à forma que ao conteúdo visto que o conteúdo dos dados minerados não apresentam em sua totalidade a realidade do cenário acadêmico científico atual.

Assim os resultados gerados a partir dos dados minerados do CiteSeerX mostraram que a maioria dos dados recuperados é referente às universidades dos Estados Unidos, e que essas universidades possuem redes de colaboração com várias universidades ao redor do mundo, mas as redes de colaboração mais comuns são entre universidades dentro do mesmo país.

Já os resultados obtidos a partir dos dados minerados do DBLP mostram que a maioria dos autores e instituições é provinda da China e Alemanha.

Os autores com maior número de artigos publicados possuem vasta rede de relacionamento e cada um deles trabalha em colaboração com vários outros autores de diferentes universidades.

Observou-se que os assuntos mais publicados em revistas e conferências são basicamente os mesmos, diferenciando-se apenas na quantidade de registros para cada tipo de evento.

Por fim, os resultados alcançados através da aplicação do modelo proposto nesta pesquisa permitem a análise e a descoberta de informações sobre as redes de relacionamento entre as universidades, os assuntos mais tratados dentro do escopo buscado e os autores mais relevantes e suas respectivas áreas de atuação.

5.1 - TRABALHOS FUTUROS

Como proposta para trabalhos futuros são sugeridas as seguintes linhas de atuação:

- A ampliação do escopo da mineração de dados aplicando o modelo elaborado em diferentes bases de dados acadêmicas;
- Filtrar a mineração de dados por um país ou continente específico, a fim de fazer um estudo focado em uma determinada região;
- Filtrar a mineração de dados por data de publicação, com o intuito de descobrir, por exemplo, os assuntos mais publicados em determinado período e a evolução dos assuntos tratados por autores, universidades, veículos de publicação.

5.2 - LIMITAÇÕES DO TRABALHO

A seguir são apresentadas as limitações deste trabalho:

- Escopo de extração de dados limitado.

  Como o objetivo principal do trabalho foi o da criação de um modelo focado em estruturar o processo de obtenção dos dados de diversas bases de dados acadêmicas e posteriormente a geração de resultados e análises a partir de tal

estrutura, o escopo de mineração de dados da presente pesquisa foi limitado e os resultados gerados não refletem a realidade do mundo acadêmico científico atual.

- Algoritmos de mineração de dados (webcrawlers) dependentes da estrutura atual dos websites.

Para mineração dos dados complementares utilizados neste trabalho foi necessária a criação de vários algoritmos para mineração desses dados, em que cada algoritmo busca um dado específico dada a URL da publicação.

Para obter o dado requerido cada algoritmo faz a leitura do código HTML contido na página informada e busca pela informação requisitada. No entanto, essa busca está atrelada à estrutura de criação das páginas HTML de cada site, sendo que se houverem alterações significativas na estrutura das páginas é possível que o algoritmo não consiga buscar mais as informações e necessite de alterações para voltar a funcionar adequadamente.

- Algoritmos de mineração de dados (*webcrawlers*) do DBLP minerando apenas em três fontes de dados.

Diferentemente das URL's das publicações providas pelo CiteSeerX em que os endereços remetem ao *website* do próprio CiteSeerX onde são contidas as informações complementares, as URL's das publicações providas pelo DBLP remetem aos mais diversos *websites* acadêmicos, tais como IEEE Xplorer, ACM Digital Library, Thomsons Web Reuters, Springer Link, entre outros.

Portanto os algoritmos criados para minerar os dados complementares do DBLP foram concebidos para interpretar e capturar as informações em apenas três fontes de publicações (IEEE Xplore, Acm Digital Library e Springer Link). Sendo que para outras fontes não foi feita a mineração dos dados complementares não contidos na estrutura original.

REFERÊNCIAS BIBLIOGRÁFICAS

Liu, L. Ozsu M. Tamer. Extraction, Transformation, and Loading. Encyclopedia of Database Systems. pp 1095-1101. 2009.

Liang S., Chien C. Method and apparatus for supervising extraction/transformation/loading processes within a database system. Taiwan Semiconductor Manufacturing Company, Ltd. 2007.

Jain, A. K., Murty, M.N., & Flynn, P. J. Data clustering: a review. ACM Computing Surveys. 31(3)264-323. 1999.

Stuart P. Lloyd, Least Squares Quantization in pcm, IEEE Transactions on Information Theory, IT-28, 2: 129–37. 1982.

Jain, A. K., & Dubes, Richard C. Algorithms for Clustering Data. Prentice Hall. Upper Saddle River, NJ, USA. 1988.

Manning C. D., Raghavan P., & Schütze H., Introduction to Information Retrieval, Cambridge University Press. 2008.

Owen, S. Anil, R. Dunning, T. Friedman, E. Mahout in Action. Manning Publications. 2012.

Fiala D. Mining citation information from CiteSeer data. Akadémiai Kiadó, Budapest, Hungary. 2010.

Zaiane O. R., Chen J., Goebel R. DBconnect: Mining Research Community on DBLP Data. WebKDD/SNA-KDD '07 Proceedings of the 9th WebKDD and 1st SNA-KDD. 2007.

Benghabrit, A. Ouhbi, B. Frikh, B. Text Document Clustering with Hybrid Feature Selection. IIWAS '13 Proceedings of International Conference on Information Integration and Web-based Applications & Services. 2013.

Sinoara, R. A. Sundermann. C. V., Marcacini R. M., Domingues M. A., Rezende S. O. (2014). Named Entities as Privileged Information for Hierarchical Text Clustering. International Database Engineering & Applications Symposium, 18th, 2014.

Choi H., Kim B., Jung Y., Choi S. Korean scholarly information analysis based on Korea Science Citation Database (KSCD). Collnet Journal of Scientometrics and Information Management. 2013.

Chikhaoui B, Chiazzaro M, Wang S. A New Granger Causal Model for Influence Evolution in Dynamic Social Networks: The Case of DBLP. AAAI Conference on Artificial Intelligence Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

Thasleena N.T, Uday Babu P, Varghese S.C. Effective XML Document Classification with Protection from Ambiguous Class Prediction Performance Evaluation with DBLP & Wikipedia Corpus. The International Conference on Information Science. 2014.

Tchuente D, Canut M, Jessel N, Peninou A, Sèdes F. A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on Facebook and DBLP. Social Network Analysis and Mining September 2013, Volume 3, Issue 3, pp 667-683. 2013.

Way S. F., Larremore D. B., Clauset A. Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks. Social and Information Networks. 2016.

Blondel V. D., Guillaume J. L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech. 2008.

Leydesdorff L., Bihui J. Mapping the Chinese Science Citation Database in Terms of Aggregated Journal–Journal Citation Relations. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 56(14):1469–1479, 2005.

Zhang L., Liuc X., Janssensb F., Lianga L., Wolfgang Glänze. Subject clustering analysis based on ISI category classification. Journal of Informetrics 4 (2010) 185–193.

Witten, I. H., Frank, E. Hall M. A. 2011. Data mining : practical machine learning tools and techniques.—3rd ed. Elsevier. 2011.

Qian G., Sural S., Gu y, & Pramanik S. Similarity between euclidean and cosine angle distance for nearest neighbor queries. In Proceedings of the 2004 ACM symposium on Applied computing, SAC '04, pages 1232–1237, New York, NY, USA, ACM. 2004.

G. Salton, M.J. McGill Introduction to Modern Information Retrieval. McGraw-Hill Book Company 1983.

Wu J., Williams K., Hung-Hsuan C., Khabsa M., Caragea C., Ororbia A., Jordan D., Giles C. L. CiteSeerX: AI in a Digital Library Search Engine. AAAI Publications, Twenty-Sixth IAAI Conference. 2014.

Borgatti, S.P. NetDraw Software for Network Visualization. Analytic Technologies: Lexington, KY. 2002.

ANEXOS

ANEXO 1 - CÓDIGO FONTE DO ALGORITMO DE EXTRAÇÃO DOS DADOS DO CITESEERX COM O USO DA API

```
private static void crawlerCiteSeerXData() {
    try {
        List<RecordVO> records = new ArrayList<RecordVO>();
        RecordVO recordVo;
        Record record;
        URL url = new URL("http://citeseerx.ist.psu.edu/oai2");
        OAIReader oaiReader = new OAIReader(url, (String) null, (String) null, (String) null, "oai_dc");
        int i = 0;
        while ((record = oaiReader.readNext()) != null) {
            recordVo = new RecordVO();
            recordVo.setIdentifier(record.getIdentifier());
            recordVo.setDatestamp(record.getDatestamp());
            recordVo.setMetadata(record.getMetadata());
            recordVo.setRecordXML(record.getRecordXML());
            recordVo.setDeleted(record.isDeleted());
            Iterator setSpecs = record.getSetSpecs();
            if (setSpecs != null) {
                while (setSpecs.hasNext()) {
                    recordVo.getSpecs().add(setSpecs.next().toString());
                }
            }
            Iterator abouts = record.getAbouts();
            if (abouts != null) {
                while (abouts.hasNext()) {
                    recordVo.getAbouts().add(abouts.next().toString());
                }
            }
            records.add(recordVo);
            if(i % 25000 == 0){
                saveFile(records,i);
                records = null;
                records = new ArrayList<RecordVO>();
            }
        }
    } catch (Exception e) {
        e.printStackTrace();
    }
}
```

ANEXO 2 - CÓDIGO FONTE DO ALGORITMO DE MINERAÇÃO DOS DADOS ADICIONAIS DO CITESEERX (Web-crawler)

```
public class CrawlerCiteseerx {

    private static void getVersionData(List<Creator> creators,
            String url,
            org.acme.oai.domain.Document document) throws IOException {
        Document doc;
        doc = Jsoup.connect(url).timeout(999999999).get();
        Elements newsHeadlines = doc.getElementsByClass("version");
        ListIterator<Element> iterator = newsHeadlines.listIterator();
        while (iterator.hasNext()) {
            Element element = iterator.next();
            Creator creator = new Creator();
            for (Node node : element.childNodes()) {
                for (Node nod : node.childNodes()) {
                    boolean autor = false;
                    boolean affiliation = false;
                    boolean address = false;
                    boolean venue = false;
                    boolean venueType = false;
                    for (Node no : nod.childNodes()) {
                        for (Node n : no.childNodes()) {
                            if (autor) {
                                creator.setCreator(n.toString());
                                autor = false;
                            } else if (affiliation) {
                                if(StringUtils.isNotBlank(n.toString().trim())){
                                    creator.setAffiliation(n.toString().trim());
                                }
                                affiliation = false;
                                creators.add(creator);
                            }else if (address) {
                                creator.setAddress(n.toString());
                                address = false;
                                creator = new Creator();
                            } else if (venue) {
                                document.setVenue(n.toString());
```

```
                            document.setVenue(n.toString());
                            venue = false;
                        }else if (venueType) {
                            document.setVenueType(n.toString());
                            venueType = false;
                        }
                        if ("AUTHOR NAME".equalsIgnoreCase(n.toString()
                                .trim())) {
                            autor = true;
                        } else if ("AUTHOR AFFIL".equalsIgnoreCase(n
                                .toString().trim())) {
                            affiliation = true;
                        } else if ("AUTHOR ADDR".equalsIgnoreCase(n
                                .toString().trim())) {
                            address = true;
                        } else if ("VENUE".equalsIgnoreCase(n.toString()
                                .trim())) {
                            venue = true;
                        } else if ("VENUE TYPE".equalsIgnoreCase(n.toString()
                                .trim())) {
                            venueType = true;
                        }
                    }
                }
            }
        }
    }

    public CrawlerCiteseerx() {
    }

    private static Integer records = 0;

    public CrawlerCiteseerx(Integer init, Integer end, String filePath) {
        List<String> identifiers = getIdentifiers(init, end);
```

```
        List<String> identifiers = getIdentifiers(init, end);
        List<String> timedOutUrls = new ArrayList<String>();
        List<Creator> creators = new ArrayList<Creator>();
        List<org.acme.oai.domain.Document> documents = new ArrayList<org.acme.oai.domain.Document>();
        org.acme.oai.domain.Document document = new org.acme.oai.domain.Document();
        for (String identifier : identifiers) {
            String url = null;
            try {
                document = new org.acme.oai.domain.Document();
                document.setIdentifier(identifier);
                url = identifier.replace("summary", "versions");
                getVersionData(creators, url, document);
                if(StringUtils.isNotBlank(document.getVenue())){
                    documents.add(document);
                }
            } catch (IOException e) {

                timedOutUrls.add(url);
                System.out.print("TIMED OUT ");
            } finally {
                System.out.println("REGISTROS: " +(records++));
            }
        }
        saveCreators(creators,filePath);
        saveDocuments(documents,filePath.replace("file_affiliation_", "file_affiliation_document_"));
    }

    private static List<String> getIdentifiers(Integer init, Integer limit) {
        DocumentDao documentDao = new DocumentDao(HibernateUtil.getSession(),
                Document.class);
        return documentDao.getAllIdentifiers(init, limit);
    }

    private static void saveCreators(List<Creator> creators, String filePath) {
        try {
```

```
            FileWriter fileWriter = new FileWriter(new File(filePath));
            BufferedWriter bufferedWriter = new BufferedWriter(fileWriter);
            for(Creator creator : creators){
                if(StringUtils.isNotBlank(creator.getAffiliation())
                        && StringUtils.isNotEmpty(creator.getAffiliation())){
                    bufferedWriter.write(creator.getCreator() + "<<SEPARATOR>>" +
                            creator.getAffiliation() + "<<SEPARATOR>>" + creator.getAddress());
                    bufferedWriter.newLine();
                }
            }
            bufferedWriter.flush();
            bufferedWriter.close();
            fileWriter.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }

    private static void saveDocuments(
            List<org.acme.oai.domain.Document> documents,String filePath) {
        try {
            FileWriter fileWriter = new FileWriter(new File(filePath));
            BufferedWriter bufferedWriter = new BufferedWriter(fileWriter);
            for(org.acme.oai.domain.Document document : documents){
                if(StringUtils.isNotBlank(document.getVenue())){
                    bufferedWriter.write(document.getIdentifier() + "<<SEPARATOR>>" +
                            document.getVenue() + "<<SEPARATOR>>" + document.getVenueType());
                    bufferedWriter.newLine();
                }
            }
            bufferedWriter.flush();
            bufferedWriter.close();
            fileWriter.close();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
```

ANEXO 3 - TRECHO DE CÓDIGO COM O MAPEAMENTO DO XML PARA A ENTIDADE DOCUMENTO EM OBJETO

```
@Entity
@Table(name="document",schema="citeseerx")
@XStreamAlias("oai_dc:dc")
public class Document implements Serializable{

    private static final long serialVersionUID = 9024915013181330136L;

    @Column(name="title",length=2000)
    @XStreamAlias("dc:title")
    private String title;

    @Transient
    @XStreamImplicit(itemFieldName="dc:creator")
    private List<String> creator;

    @ManyToMany(cascade=CascadeType.ALL)
    @JoinTable(name="document_creator", schema="citeseerx",
    joinColumns = @JoinColumn (name = "identifier"),
    inverseJoinColumns = @JoinColumn (name = "creator"))
    private List<Creator> creators;

    @ManyToMany(cascade=CascadeType.ALL)
    @JoinTable(name="document_subject", schema="citeseerx",
    joinColumns = @JoinColumn (name = "identifier"),
    inverseJoinColumns = @JoinColumn (name = "subject"))
    private List<Subject> listSubjects;

    @Transient
    @XStreamImplicit(itemFieldName="dc:subject")
    private List<String> subjects;

    @Column(name="description",length=2000)
    @XStreamAlias("dc:description")
    private String description;

    @Column(name="contributor",length=2000)
    @XStreamAlias("dc:contributor")
    private String contributor;

    @Column(name="publisher",length=2000)
```

ANEXO 4 - CONSULTA CLUSTER TIPO DE EVENTO, ASSUNTO E FILIAÇÃO

```
--sql venue-type (journal, conference) X numero registro X afiliação (top 10) X autor
select doc.venue_type, c.affiliation_address, c.creator
from citeseerx.document doc
join citeseerx.document_creator dc on dc.identifier = doc.identifier
join citeseerx.creator c on c.creator = dc.creator
where upper(trim(doc.venue_type)) in ('CONFERENCE','JOURNAL')
and
 upper(c.affiliation_address) in
 ('ENGINEERING DESIGN RESEARCH; LABORATORY; CALIFORNIA INSTITUTE OF TECHNOLOGY - PASADENA, CALIFORNIA 91125',
'SCHOOL OF COMPUTER SCIENCE; UNIVERSITY OF BIRMINGHAM - EDGBASTON, B15 2TT, UK',
'SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 15213',
'STRAIGHT-LINE COMPUTATION OF THE - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE',
'CENTRE FOR WIRELESS COMMUNICATIONS; DEPARTMENT OF ELECTRICAL &AMP; COMPUTER ENGINEERING;
UNIVERSITY OF WATERLOO, WATERLOO, ONTARIO - CANADA, N2L 3G1',
'UNIVERSIDADE DOS AÇORES - APARTADO 1422, 9501-801 PONTA DELGADA, PORTUGAL',
'ZURICH, SCHWEIZ - ECOLE SUISSE POLYTECHNIQUE FEDERALE DE ZURICH',
'DREYER URGENT CARE CENTER (DREYER UCC) IS LOCATED IN - AURORA, ILLINOIS. THE FACILITY HAS NO EMERGENCY ROOM; DOCTORS',
'DEFENSE MODELING AND SIMULATION OFFICE - 1901 N. BEAUREGARD STREET; ALEXANDRIA, VA 22311',
'FOOD CONSUMPTION AND NUTRITION DIVISION; INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE - 2033 K STREET, N.W.;
WASHINGTON, D.C. 20006 U.S.A.')
```

ANEXO 5 - CONSULTA CLUSTER TIPO DE EVENTO, ASSUNTO, FILIAÇÃO E AUTOR

```
-- sql cluster subject (top 40) X venue-type (journal, conference) X numero registro X afiliação (top 100) X autor
select doc.venue_type, upper(c.affiliation_address), upper(c.creator), upper(ds.subject)
from citeseerx.document doc
join citeseerx.document_creator dc on dc.identifier = doc.identifier
join citeseerx.document_subject ds on ds.identifier = doc.identifier
join citeseerx.creator c on c.creator = dc.creator
where upper(trim(doc.venue_type)) in ('CONFERENCE','JOURNAL')
and upper(c.affiliation_address) in
('ENGINEERING DESIGN RESEARCH; LABORATORY; CALIFORNIA INSTITUTE OF TECHNOLOGY - PASADENA, CALIFORNIA 91125',
'SCHOOL OF COMPUTER SCIENCE; UNIVERSITY OF BIRMINGHAM - EDGBASTON, B15 2TT, UK',
'SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 15213',
'STRAIGHT-LINE COMPUTATION OF THE - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE',
'CENTRE FOR WIRELESS COMMUNICATIONS; DEPARTMENT OF ELECTRICAL &AMP; COMPUTER ENGINEERING; UNIVERSITY OF WATERLOO, WATERLOO, ONT
'UNIVERSIDADE DOS AÇORES - APARTADO 1422, 9501-801 PONTA DELGADA, PORTUGAL',
'ZURICH, SCHWEIZ - ECOLE SUISSE POLYTECHNIQUE FEDERALE DE ZURICH',
'DREYER URGENT CARE CENTER (DREYER UCC) IS LOCATED IN - AURORA, ILLINOIS. THE FACILITY HAS NO EMERGENCY ROOM; DOCTORS',
'DEFENSE MODELING AND SIMULATION OFFICE - 1901 N. BEAUREGARD STREET; ALEXANDRIA, VA 22311',
'FOOD CONSUMPTION AND NUTRITION DIVISION; INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE - 2033 K STREET, N.W.; WASHINGTON, D.C.
'ECOLE NORMALE SUPÉRIEURE DE LYON - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE',
'PROCEEDINGS FORMATTING TEAM - JOHN W. LOCKHART, RED HAT, INC.',
'1 IICM- SOFTWARETECHNOLOGY, GRAZ UNIVERSITY OF TECHNOLOGY, - INFFELDGASSE 16B/I A-8010 GRAZ, AUSTRIA',
'MODEL VERIFICATION AND VALIDATION - JOHN S. CARSON, II',
'FACULTY OF ELECTRICAL ENG. &AMP; INFORMATION TECHNOLOGY; SLOVAK UNIVERSITY OF TECHNOLOGY - ILKOVIC?VA 3, 812 19 BRATISLAVA, SL
'TEL-AVIV UNIVERSITY, TEL-AVIV, ISRAEL, AND INSTITUTE FOR ADVANCED STUDY, PRINCETON - NEW JERSEY',
'CWI - P.O. BOX 94079, 1090 GB AMSTERDAM, THE NETHERLANDS',
'1; DEPARTMENT OF COMPUTER SCIENCE; 2; SCHOOL OF COMPUTING SCIENCE; UNIVERSITY OF MANCHESTER; UNIVERSITY OF NEWCASTLE UPON TYNE
'ENVIRONMENT AND PRODUCTION TECHNOLOGY DIVISION; INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE - 2033 K STREET, N.W.; WASHINGTON
'MOTOR VEHICLES; SIMULATION AS A PRIMARY TOOL WAS USED TO EVALUATE THE; EFFECTIVENESS OF THE SANTA TERESA DEPARTMENT OF MOTOR;
'DARMSTADT UNIVERSITY OF TECHNOLOGY, DEPT. OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY; INDUSTRIAL PROCESS &AMP; SYSTE
'; NOTES; © 2005 COLD SPRING HARBOR LABORATORY PRESS; &AMP;QUOT;SUPPLEMENTAL RESEARCH DATA&AMP;QUOT; - ; THIS ARTICLE CITES 27
'† DEPARTMENT OF ELECTRICAL ENGINEERING &AMP; COMPUTER SCIENCE, UNIVERSITY OF MICHIGAN, ANN ARBOR - MI 48109',
'DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BOLOGNA - ITALY',
'DISTRIBUTED SYSTEMS LABORATORY; DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE, UNIVERSITY OF PENNSYLVANIA; 200 S. 33RD STREET
'COMPUTER ENGINEERING LABORATORY,; ELECTRICAL ENGINEERING DEPARTMENT,; DELFT UNIVERSITY OF TECHNOLOGY, - DELFT, THE NETHERLANDS
'DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE; ODENSE UNIVERSITY - DENMARK',
'DEPARTMENT OF INDUSTRIAL &AMP; MANUFACTURING ENGINEERING - 310 LEONHARD BUILDING',
'DEPARTMENT OF COMPUTER SCIENCE; HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY - CLEAR WATER BAY, KOWLOON, HONG KONG',
'INTELLIGENT SYSTEMS INSTITUTE,; NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST) - AIST TSUKUBA CENTRAL
'; IMAGE-CENTRIC DATABASES; COMMENTARY - ; 00101011100101; 10010100111010; 10101000101001',
'COMPUTER SCIENCE DEPARTMENT; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 15213 USA',
'SCHOOL OF COMPUTER SCIENCE AND IT; UNIVERSITY OF NOTTINGHAM - NOTTINGHAM NG8 1BB, UK',
'DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY; UNIVERSITY OF TRENTO- ITALY - VIA SOMMARIVE 14, 38050 POVO TRENTO',
'DIPARTIMENTO DI INFORMATICA E SISTEMISTICA; UNIVERSITÀ DI ROMA "LA SAPIENZA" - VIA SALARIA 113, I-00198 ROMA, ITALY',
'1 DIPARTIMENTO DI SCIENZE DELL'INFORMAZIONE, UNIVERSITÀ DEGLI STUDI DI MILANO (ITALY) - 2 LORIA &AMP; INRIA-LORRAINE',
'UC BERKELEY, EECS DEPARTMENT, BERKELEY - CA 94720-1770, U.S.A.',
'COLLEGE OF COMPUTING; GEORGIA INSTITUTE OF TECHNOLOGY - ATLANTA, GA 30332',
'2 INSTITUTE FOR SOFTWARE INTEGRATED SYSTEMS, VANDERBILT UNIVERSITY - NASHVILLE TN 37235',
'IRST - 38050 POVO (TN), ITALY, DISA, VIA INAMA 5, 38100 TRENTO, ITALY.;',
'UNIVERSITY OF MINNESOTA, DEPARTMENT OF COMPUTER SCIENCE - MINNEAPOLIS, MN 55455',
'DIVISION OF ENGINEERING - P.O. BOX 111',
'DEPARTMENT OF COMPUTER SCIENCE, YALE UNIVERSITY - NEW HAVEN, CT 06520-8285, U.S.A.',
'EMBEDDED AND MOBILE COMPUTING CENTER (EMC 2); THE PENNSYLVANIA STATE UNIVERSITY - UNIVERSITY PARK, PA, 16802, USA',
'UNIVERSITY OF DORTMUND - GERMANY',
'ICIR/ICSI, BERKELEY - CA 94704, USA',
'VIRGINIA IMAGE AND VIDEO ANALYSIS (VIVA); DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING,; UNIVERSITY OF VIRGINIA, CHARLOTT
'ABSTRACT; DEPARTMENT OF COMPUTER SCIENCE DEPARTMENT OF INFORMATION SCIENCE; UNIVERSITY OF OTAGO - NEW ZEALAND UNIVERSITY OF OT
'? DEPARTMENT OF COMPUTER SCIENCE; ILLINOIS INSTITUTE OF TECHNOLOGY - 10 W. 31ST STREET, CHICAGO, IL 60616',
'2 COMPLUTENSE UNIVERSITY OF MADRID; DEPARTMENT OF SOFTWARE ENGINEERING AND ARTIFICIAL INTELLIGENCE; PROFESOR JOSÉ GARCIA SANTE
'LSV- CNRS &AMP; ENS DE CACHAN - 61, AVENUE DU PRÉSIDENT WILSON, 94230 CACHAN, FRANCE'
)
and upper(ds.subject) in
('DESIGN','ALGORITHMS','PERFORMANCE','CONTENTS',
'EXPERIMENTATION','SECURITY','GENERAL TERMS','CATEGORIES AND SUBJECT DESCRIPTORS','MEASUREMENT',
'SIMULATION','DATA MINING','LANGUAGES','HUMAN FACTORS','CLASSIFICATION','INFORMATION RETRIEVAL','NEURAL NETWORKS',
'RELIABILITY','XML','CLUSTERING','ONTOLOGY','OPTIMIZATION','SEMANTIC WEB','THEORY',
'SCHEDULING','MACHINE LEARNING','MANAGEMENT','INTERNET','GENETIC ALGORITHMS','JAVA',
'VERIFICATION','VISUALIZATION','EVALUATION','GENERAL TERMS ALGORITHMS','VIRTUAL REALITY',
'ROUTING','WEB SERVICES','DISTRIBUTED SYSTEMS','UML','LEARNING','CATEGORY')
```

ANEXO 6 - CONSULTA RANKING TIPO DE EVENTO

```
-- top venue-type
select venue_type, sum(identifier) as qtd from (
select upper(venue_type) venue_type, count(identifier) as identifier
from citeseerx.document
where venue_type is not null and venue_type <> ''
group by venue_type
order by 2 desc) as tem
group by venue_type
order by 2 desc
```

ANEXO 7 - CONSULTA RANKING ASSUNTO

```
-- sql top subject
select subject, sum(identifier)from (
select upper(s.subject) subject, count(ds.identifier) as identifier
from citeseerx.subject s
join citeseerx.document_subject ds on ds.subject = s.subject
where s.subject <> ''
group by s.subject
order by 2 desc) as tem
group by  subject
order by 2 desc
```

ANEXO 8 - CONSULTA RANKING FILIAÇÃO

```
-- sql top affiliation
select affiliation, sum(identifier) as qtd from (
select upper(c.affiliation) affiliation, count(dc.identifier) as identifier
from citeseerx.creator c
join citeseerx.document_creator dc on c.creator = dc.creator
where c.affiliation is not null and c.affiliation <> ''
group by c.affiliation
order by 2 desc) as tem
group by affiliation
order by 2 desc
```

ANEXO 9 - CONSULTA RANKING AUTOR

```
--top creator
select creator, sum(identifier) as qtd from (
select upper(c.creator) creator, count(dc.identifier) as identifier
from citeseerx.creator c
join citeseerx.document_creator dc on c.creator = dc.creator
where c.creator is not null and c.creator <> ''
group by c.creator
order by 2 desc) as tem
group by creator
order by 2 desc
```

ANEXO 10 - CONSULTA RANKING INSTITUIÇÕES POR TIPO DE EVENTO

```
--top 5 institutes per venue type
select upper(doc.venue_type) as venue_type, upper(c.affiliation_address) subject, count(doc.identifier) as identifier
from citeseerx.document doc
join citeseerx.document_creator dc on dc.identifier = doc.identifier
join citeseerx.creator c on dc.creator = c.creator
where upper(trim(c.affiliation_address)) in
('ENGINEERING DESIGN RESEARCH; LABORATORY; CALIFORNIA INSTITUTE OF TECHNOLOGY - PASADENA, CALIFORNIA 91125',
'SCHOOL OF COMPUTER SCIENCE; UNIVERSITY OF BIRMINGHAM - EDGBASTON, B15 2TT, UK',
'SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 15213',
'STRAIGHT-LINE COMPUTATION OF THE - 46 ALLÉE D'ITALIE, 69364 LYON CEDEX 07, FRANCE',
'CENTRE FOR WIRELESS COMMUNICATIONS; DEPARTMENT OF ELECTRICAL; COMPUTER ENGINEERING; UNIVERSITY OF WATERLOO',
'UNIVERSIDADE DOS AÇORES - APARTADO 1422, 9501-801 PONTA DELGADA, PORTUGAL',
'ZURICH, SCHWEIZ - ECOLE SUISSE POLYTECHNIQUE FEDERALE DE ZURICH')
and upper(trim(doc.venue_type)) in ('CONFERENCE','JOURNAL')
group by upper(venue_type), upper(c.affiliation_address)
order by 3 desc
```

ANEXO 11 - CONSULTA RANKING ASSUNTO POR TIPO DE EVENTO

```
--TOP 10 Subjects per venue type
select upper(doc.venue_type) as venue_type, upper(s.subject) subject, count(doc.identifier) as identifier
from citeseerx.subject s
join citeseerx.document_subject ds on ds.subject = s.subject
join citeseerx.document doc on doc.identifier = ds.identifier
where upper(s.subject) in
('KEY WORDS','DESIGN','ALGORITHMS','PERFORMANCE','CONTENTS','EXPERIMENTATION',
'SECURITY','GENERAL TERMS','CATEGORIES AND SUBJECT DESCRIPTORS','MEASUREMENT','SIMULATION')
and upper(trim(doc.venue_type)) in ('CONFERENCE','JOURNAL')
group by upper(venue_type), upper(s.subject)
order by 3 desc
```

ANEXO 12 - CONSULTA CRIAÇÃO REDE DE RELACIONAMENTO

```
-- author - author - affiliation - affiliation
select distinct upper(trim(c.creator)) creator, c.affiliation_address, c1.creator, c1.affiliation_address
from citeseerx.creator c
join citeseerx.document_creator dc on dc.creator = c.creator
join citeseerx.document doc on doc.identifier = dc.identifier
join citeseerx.document_creator dc1 on dc1.identifier = doc.identifier
join citeseerx.creator c1 on dc1.creator = c1.creator
where upper(trim(c.creator)) in
('J. M. CHARNES','J. L. SNOWDON','E. YÜCESAN','NOGA ALON','C. -H. CHEN',
'J. A. JOINES','F. B. ARMSTRONG','N. M. STEIGER','M. E. KUHL','B. A. PETERS','J. S. SMITH')
and c.affiliation_address <> '' and c1.affiliation_address <> ''
order by 1,3
```

ANEXO 13 - ALGORITMO PARA GERAÇÃO DA REDE DE RELACIONAMENTO DAS FILIAÇÕES

```
protected void createMatrix(String path){
    populateAffiliationSubject(25, 20);
    WritableWorkbook workbook = null;

    try {
        workbook = Workbook.createWorkbook(new File(path));
        WritableSheet sheet = workbook.createSheet("creator", 0);
        sheet.addCell(new Label(0, 0, "Id"));
        int i = 1;
        for(String affiliation : topAffiliation){
            sheet.addCell(new Label(0, i, affiliation));
            i++;
        }
        int j = 1;
        for(String affiliation : affiliations){
            sheet.addCell(new Label(j, 0, affiliation));
            j++;
        }
        int r = 1;
        for(AffiliationColaboration creatorColaboration : getAffiliationColaborations()){
            int c = 1;
            for(String affiliation : affiliations){
                if(creatorColaboration.getColaborator().contains(affiliation)){
                    sheet.addCell(new Label(c, r, "1"));
                }else{
                    sheet.addCell(new Label(c, r, "0"));
                }
                c++;
            }
            r++;
        }

        workbook.write();
        workbook.close();

    } catch (Exception e) {
        e.printStackTrace();
    }
}
```

ANEXO 14 - TRECHO DO ARQUIVO ARFF CRIADO PARA GERAÇÃO DOS CLUSTERS

```
@relation venue_affiliation_subject_author
@attribute venue_type {'JOURNAL','CONFERENCE'}
@attribute affiliation {'; 2; 1; GRADUATE SCHOOL OF ENGINEERING SCIENCE, OSAKA UNIVERSITY;
@attribute author {'ADITYA AKELLA','ALBERT T. CORBETT','ALBERTO MONTRESOR','ANGELA Z. WAGN
@attribute subject {'DESIGN','ALGORITHMS','PERFORMANCE','CONTENTS','EXPERIMENTATION','SECU
@data
'CONFERENCE','INSTITUTE FOR INFOCOMM RESEARCH - 21 HENG MUI KENG TERRACE, SINGAPORE 119613
'CONFERENCE','INSTITUTE FOR INFOCOMM RESEARCH - 21 HENG MUI KENG TERRACE, SINGAPORE 119613
'CONFERENCE','SCHOOL OF COMPUTER SCIENCE; CARNEGIE MELLON UNIVERSITY - PITTSBURGH, PA 1521
'CONFERENCE','THEME NUM--- SYSTEMES NUMERIQUES - 655, AVENUE DE LEUROPE, 38334 MONTBONNOT
'CONFERENCE','DARMSTADT UNIVERSITY OF TECHNOLOGY, DEPT. OF ELECTRICAL ENGINEERING AND INFO
'JOURNAL','UNIVERSITY OF MINNESOTA, DEPARTMENT OF COMPUTER SCIENCE - MINNEAPOLIS, MN 55455
'JOURNAL','UNIVERSITY OF MINNESOTA, DEPARTMENT OF COMPUTER SCIENCE - MINNEAPOLIS, MN 55455
'CONFERENCE','DISTRIBUTED SYSTEMS LABORATORY; DEPARTMENT OF COMPUTER AND INFORMATION SCIEN
'CONFERENCE','DISTRIBUTED SYSTEMS LABORATORY; DEPARTMENT OF COMPUTER AND INFORMATION SCIEN
'CONFERENCE','MAX-PLANCK-INSTITUT FÜR INFORMATIK - STUHLSATZENHAUSWEG 85; D-66123, SAARBRÜ
'CONFERENCE','UNIVERSITY OF MINNESOTA, DEPARTMENT OF COMPUTER SCIENCE - MINNEAPOLIS, MN 55
'JOURNAL','ABSTRACT; 2 FACULTY OF COMPUTER SCIENCE; FREE UNIVERSITY OF BOLZANO/BOZEN - PIA
'JOURNAL','DIPARTIMENTO DI INFORMATICA E SISTEMISTICA; UNIVERSITÀ DI ROMA "LA SAPIENZA" -
'CONFERENCE','ICIR/ICSI, BERKELEY - CA 94704, USA','SCOTT SHENKER','DESIGN'
'CONFERENCE','ICIR/ICSI, BERKELEY - CA 94704, USA','SCOTT SHENKER','PERFORMANCE'
'JOURNAL','; NOTES; © 2005 COLD SPRING HARBOR LABORATORY PRESS; &AMP;QUOT;SUPPLEMENTAL RES
'JOURNAL','DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BOLOGNA - ITALY','OZALP BABAOGLU'
'JOURNAL','DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF BOLOGNA - ITALY','ALBERTO MONTRES
```