Um método de indexação e organização de informações baseado em um sociograma – Aplicação em domínios específicos

Rafael Loureiro S. de Moraes, Jonice de Oliveira Sampaio

Programa de Pós-graduação em Informática – Universidade Federal do Rio de Janeiro (PPGI/UFRJ) – Rio de Janeiro, RJ – Brasil

{rafaeloureirosm, jonice}@ufrj.br

Abstract. *The growth of bibliographic production in all areas of science is evident, but the organization of its information is still a challenge. In this way the work starts from a proposal to improve an information organization, in this case a taxonomy, using the analysis of social networks to assist in the management and updating of knowledge. Therefore, we have developed a methodology ranging from the choice of publications, definition of a reference taxonomy, application of social network analysis metrics to the recommendation of a new taxonomy. We evaluated the method quantitatively and qualitatively, where we obtained consistent results and were able to recommend new ways of organizing information*

Resumo. *O crescimento da produção bibliográfica em todas as áreas da ciência é evidente, porém a organização de suas informações ainda é um desafio. Diante disso o trabalho parte de uma proposta de criar um método de análise sociométrica buscando melhorar uma organização da informação, neste caso uma taxonomia, utilizando a análise de redes sociais para auxiliar na gestão e atualização do conhecimento. Sendo assim, desenvolvemos uma metodologia que vai desde a escolha das publicações, definição de uma taxonomia de referência, aplicação de métricas de análise de redes sociais até a recomendação de uma nova taxonomia. Avaliamos o método quantitativamente e qualitativamente, onde obtivemos resultados consistentes e conseguimos recomendar novas formas de organização da informação.*

1. Introdução

O ‘caos’ documentário e a ‘explosão’ da informação trouxeram mudanças significativas na trajetória da comunicação do conhecimento (SOUZA, R. 2006). Os processos de produção, tratamento, disseminação e, principalmente, organização das informações tornaram-se crucias para pesquisadores e instituições de ensino. A organização do conhecimento é importante para a gestão e avaliação, as agências de fomento em Ciência e Tecnologia (C&T) utilizam essas informações como um dos suportes para o desenvolvimento de políticas públicas, lidam com a categorização do conhecimento como elemento de referência para ensino, pesquisa e inovação.

Para (CASTRO, 1985) a produção cientifica é algo tangível, que pode ser contado e avaliado, sendo assim avaliar sua produção científica em uma determinada área de atuação ou evento, é mensurar a produção cientifica naquela área ou evento.

Outra base de dados muito conhecida e importante é a DBLP que é um repositório bibliográfico de ciência da computação, na imagem a seguir podemos verificar um crescimento nos registros e publicações, reforçando ainda mais o que comentamos no parágrafo anterior que mostra um crescimento da produção científica.

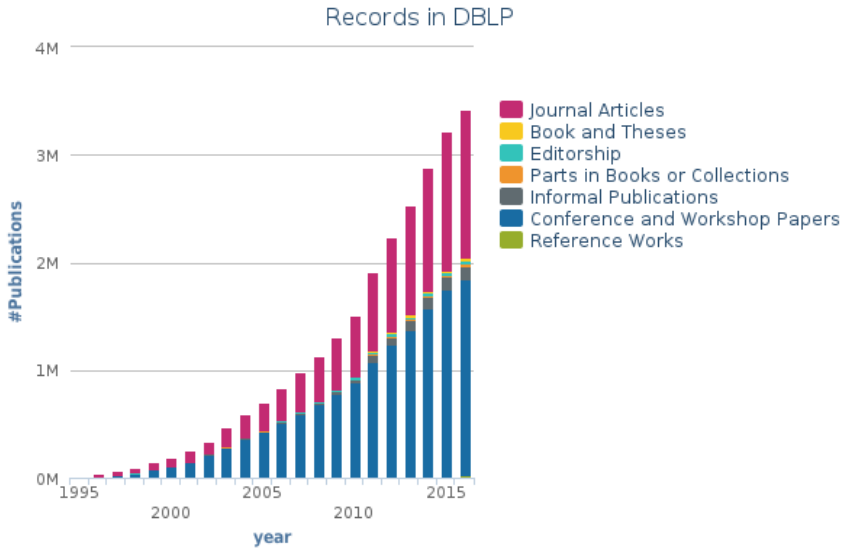

Figura 1 - Crescimento da produção acadêmica mundial (Fonte: DBLP - 14/07/2016)

Devido a todo esse crescimento mundial (Figura 1) é imprescindível saber organizar todas essas informações. Pois com o crescimento das produções acadêmicas, crescem juntamente a procura por artigos, revistas, conferências, áreas de atuação em destaque, pesquisadores de referência etc. E a recuperação dessas informações é realizada através de buscas em bibliotecas virtuais, bases indexadas, onde grandes partes das buscas são realizadas por meio de palavras-chave.

Uma palavra-chave pode ser entendida como a menor unidade de um ou mais termos que sintetiza e identifica o conteúdo de um documento de texto completo permitindo, assim, uma representação simplificada desse documento ou podendo servir como referência em uma pesquisa (KAUR; GUPTA, 2010; ROSE et al., 2010). A palavra-chave também pode ser entendida como uma unidade que sintetiza e compacta uma área de conhecimento em termos.

As áreas de conhecimento podem ser entendidas como áreas que se destacam em uma determinada ciência, onde vários pesquisadores trabalham e desenvolvem pesquisas baseadas naquele tema. Palavras-chave são candidatas a se tornarem uma área de atuação mediante análise da frequência de sua utilização durante um período de tempo, e a relação delas no contexto já organizado das áreas de atuação ou áreas de conhecimento.

As classificações das áreas de atuação são determinadas por processos de organização da informação, desde a criação de dicionários controlados, taxonomias,

tesauros, podendo chegar a uma complexa ontologia. Em tempos de Big Data e computadores mais velozes, as elaborações desses processos de organização ganham fortes aliados, além do suporte de especialistas das áreas para criação desses artefatos organizacionais. A análise de todos esses dados juntamente com uma visão dos especialistas das áreas pode-se chegar a uma organização da informação mais atual e em outros casos podemos encontrar divergências ou relacionamentos entre áreas.

De acordo com (SOUZA, R. 2006), o Brasil conta com poucos instrumentos especialmente criados e desenvolvidos especificamente para a classificação de dados provenientes de atividades de pesquisa e ensino em ciência. O instrumento mais conhecido, que por tempo muito foi o único, e que ainda hoje é reconhecido como referencial por ser o mais utilizado pelas diversas instituições do país é a tabela conhecida como "tabela de áreas do conhecimento do CNPq", cuja estrutura de base data da década de 50 e sua última versão é do ano de 2012.

O IEEE (*Institute of Electrical and Electronics Engineers*) conta com uma taxonomia própria, elaborada a partir do Tesauro também elaborado pelo instituto. O IEEE Thesaurus Terms é um vocabulário controlado com mais de 9.700 termos técnicos e científicos que proporcionam, assim, um vocabulário controlado de áreas para ajudar à categorizar ou procurar conceitos de engenharia e computação. A última versão desses documentos é 2014.

A ACM (*Association for Computing Machinery*) também conta com um sistema de classificação chamado de ACM *Computing Classification System* que é um sistema de classificação para as áreas da computação criado pela *Association for Computing Machinery*. O sistema é comparável com a *Mathematics Subject Classification* (MSC) no escopo, objetivos e estrutura, sendo usado também pelas várias revistas da ACM para organizar assuntos por área. Sua última versão é de 2012.

Apesar dos avanços tecnológicos, esses processos de organização da informação são custosos, pois além de uma coleta e análise de dados, existe toda uma análise subjetiva dos especialistas de áreas especificas para avaliar e validar as informações.

Com a quantidade de informações que temos atualmente e com a velocidade que as áreas de conhecimento, principalmente da computação, vem evoluindo é importante que a organização dessas informações também acompanhe essa evolução. É notório que propor a criação ou atualização de um instrumento que organize as informações, por exemplo uma taxonomia, é uma atividade complexa onde é necessário planejamento, pessoal especializado e tempo.

O estudo da análise de redes sociais está ligado com a teoria social e com teorias matemáticas que dão suporte a pesquisa das relações entre objetos em um determinado grupo. Esses estudos estiveram sidos influenciados conceitualmente por uma visão estruturalista da sociedade (WASSERMAN E FAUST, 1994), que buscava entender a sociedade a partir da estrutura formada pelas relações entre seus indivíduos.

A dinâmica de redes considera que os indivíduos são entidades que evoluem ao longo do tempo. Suas propriedades, atributos e ligações mudam, assim como a maneira que irão interagir. Essa perspectiva nos permite entender as redes sociais como sistemas dinâmicos, não considerando apenas suas propriedades topológicas/estruturais, mas também suas propriedades dinâmicas.

Um desses aspectos de estudo é analisar as áreas de atuação dos pesquisadores em uma rede social, por exemplo uma rede de coautoria. Isso pode refletir na construção de novos conhecimentos ou de novas ligações entre áreas, até então, distintas e por isso, merecem especial atenção no que se refere à sua análise, evolução e avaliação.

2. Trabalhos Relacionados

Esse capítulo tem o objetivo demostrar trabalhos relacionados com o tema do trabalho, fazendo uma leitura crítica sobre eles e extraindo o conhecimento necessário para aplicar neste trabalho.

A leitura crítica de trabalhos científicos não deve ser encarada apenas como aprendizado. O pesquisador deve exercer, antes de tudo, o espírito crítico, para questionar a validade de todas as informações registradas nos textos.

Algumas perguntas foram realizadas para apoiar essa leitura crítica. Foram elas:

a) *De onde o autor parece tirar suas ideias?*
b) *O que foi obtido como resultado desse trabalho?*
c) *Como esse trabalho se relaciona com outros na mesma área?*
d) *Qual seria o próximo passo razoável para dar continuidade a essa pesquisa?*
e) *Como esse trabalho se relaciona com o nosso trabalho?*

Dividimos os trabalhos relacionados em 4 partes: organização do conhecimento e classificação, extração de informações e criação de grafos, estudos sociométricos em grafos e a gestão do conhecimento e redes sociais. Todas as seções seguintes são apresentadas com uma justificativa de escolha desses temas.

Organização do Conhecimento e Classificação

Os trabalhos de (ALMEIDA CAMPOS, M. L., e GOMES, H. E.), (OLIVEIRA, D., et. al, 2013), (SOUZA, R. F, 2006, 2006b) discutem a importância da classificação para as instituições de ensino e pesquisa e também para a portais institucionais, bibliotecas digitais, além de melhorar a recuperação das informações.

Destacamos dois trabalhos de (SOUZA, R. F, 2006, 2006b) que analisam exemplos de tabelas e esquemas de classificação em ciência e tecnologia com o objetivo de identificar agregações de áreas de conhecimento em diferentes necessidades de produção e uso de informação. A autora mostra que analisando esquemas de classificação bibliográfica, tabelas de classificação de áreas de conhecimento para propósitos de comunicação em ciência, administração de programas e agências de fomento, existe um consenso na agregação de áreas em grandes áreas, embora ocorram diferenças no número e na ordem de apresentação das grandes áreas em função da natureza do objeto de representação, assim como da finalidade da organização do conhecimento. No seu segundo trabalho, seu objetivo e contribuir com o debate para a revisão da tabela de áreas de conhecimento e a importância de sua atualização para a comunidade cientificam.

Atualmente boa parte dos trabalhos nacionais sobre o assunto organização do conhecimento são produzidos pelo ISKO-Brasil. A ISKO é uma sociedade fundada em 1989 que se tornou o principal grupo de estudos nessa área. O capítulo brasileiro foi

fundado em 2007 e desde então vem promovendo trabalhos e pesquisas que proporcionam o desenvolvimento do conhecimento na área (J. Guimarães e V. Dodebei, 2012; 2013; 2015). Os trabalhos descritos são publicados pelo ISKO-Brasil como uma série de estudos avançados em organização do conhecimento e formam hoje uma das melhores reuniões de trabalhos nesta área.

Extração de Informações e criação de Grafos

No trabalho apresentado por (ABILHOA, W. D., 2014) o twitter é utilizado como ferramenta de extração de palavras para a criação de grafos. Ele apresenta um método chamado TKG (Twitter Keyword Graph), onde faz a extração de palavras-chave de coleções de tweets e aplica medidas de centralidade para encontrar nós relevantes. Ele faz a comparação do seu método com o TF-IDF e KEA , e de acordo com o autor, os experimentos realizados mostraram superiores, em alguns casos, os algoritmos comparados.

Outro exemplo foi realizado por (DIAS, T. M. R., e MOITA, G. F., 2014) que extraem as palavras-chave das publicações da Plataforma Lattes para construir um grafo com elas, com intuito de identificar as mais relevantes e, ainda, aquelas que têm maior impacto na rede. Para isso foi utilizado métricas de análise de redes sociais para identificação das palavras em destaque. Utilizaram os currículos de pesquisados do grupo de pesquisadores do Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, onde foi possível observar as palavras utilizadas em conjunto com maior frequência dentro do grupo. Eles concluem que a análise de palavras-chave de publicações científicas pode ajudar na compreensão na evolução de temas de pesquisa e como eles se comportam.

Estudos Sociométricos em Grafos

O trabalho de (BARBOSA D. et. al, 2011) apresenta uma análise de uma rede de coautoria, formada por pesquisadores reais extraídos da plataforma Lattes. A análise proposta foi na determinação de vértices mais importantes, aplicando as medidas de centralidade: grau, proximidade, intermediação, autovetor e pagerank. Também foram aplicados sobre os seus dados os algoritmos clássicos de estudos sobre comunidades como o de Girvam e Newman, por exemplo. Com esse estudo os autores consideram que com os resultados obtidos podem determinar os principais autores, grupos e os principais assuntos pesquisados no Brasil

No trabalho de (FERNANDA, P., 2013), ela faz um estudo que se utiliza de técnicas de bibliometria para verificar o comportamento dos pesquisadores e as suas áreas de atuação. Ela utilizou trabalhos publicados no ONTOBRAS nas edições de 2010, 2011 e 2012, onde realizou um estudo de coautoria, cocitação e acoplamento bibliográfico. Através dos resultados obtidos foi possível verificar o comportamento dos pesquisadores e as principais áreas atuantes no evento. Concluindo que existe uma aproximação entre áreas distintas por meio da colaboração, quando os autores de áreas diferentes escrevem juntos, e também quando citam autores de outras áreas.

Gestão do Conhecimento e Redes Sociais

De acordo com o trabalho proposto por (AZEVEDO, T. B, 2013), que pretende descrever os principais conceitos relacionados à criação do conhecimento através da análise de redes sociais, o crescimento do número de indivíduos e de organizações

interligado em redes, o aprimoramento e aplicação de técnicas, tornou-se indispensável para entendimento dessa rede. A proposta do trabalho é demonstrar teoricamente como a utilização de redes sociais pode potencializar a criação e disseminação do conhecimento.

Um estudo exploratório na literatura cientifica sobre a relação entre Ciência da Informação e a Gestão do Conhecimento sob a perspectiva das redes sociais é o trabalho realizado por (FREITAS, J. L, 2012). Utilizam nesse trabalho a base BRAPCI (Base Referencial de Artigos de Periódicos em Ciência da Informação) e analisam 50 artigos, dos quais 33 utilizam as redes como aporte de pesquisa, confirmando a ênfase dos estudos de redes sociais. O trabalho conclui que as redes sociais têm contribuído significantemente no ambiente organizacional de gestão colaborativa. Observa-se que a partir de 2006 os estudos sobre análise de redes sociais foram muito expressivos, assim como os estudos empíricos que relacionam as redes e a gestão do conhecimento. Outros trabalhos também fizeram um estudo bibliográfico utilizando ARS para analisar cursos de pós-graduação (BUFREM, L et al. 2011 e DOBEDEI, V. 2012) ou na utilização de folksonomias (MOURA, M. 2009; ASSIS, J E MOURA, M, 2013) e ontologias (MOURA, M, 2011)

3. Metodologia

A estrutura metodológica inicial para organização terminológica pautou-se na estruturação geral de taxonomias, tal qual sugere a Norma ANSI/NISO Z39.19-2005, mas com incorporações de alguns aspectos metodológicos adaptados as especificidades do trabalho.

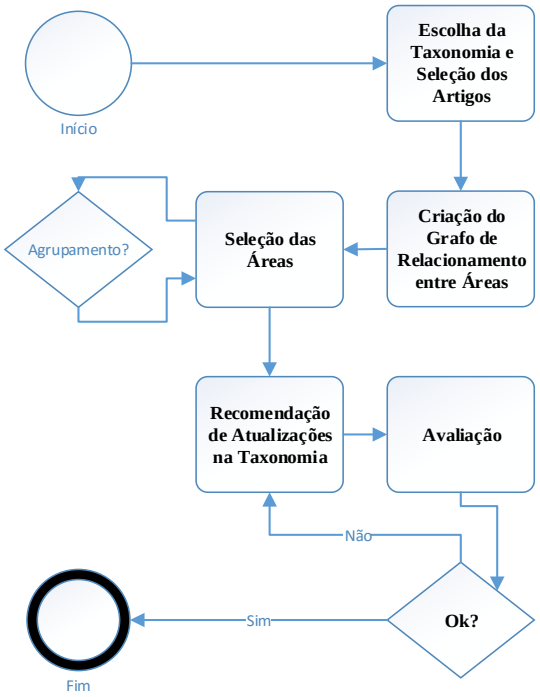

Figura 2 - Metodologia

Escolha da Taxonomia e Seleção dos Artigos

Nesta seção nosso objetivo será definir uma taxonomia como referência e selecionaremos os artigos que irão compor a nossa base de informações. Existem diversas taxonomias para organizar as informações em diferentes áreas de atuação, os exemplos mais comuns vêm da área da saúde, biologia e biblioteconomia.

Na computação alguns órgãos têm classificação própria de áreas de atuação. Podemos citar como exemplos a *The ACM Computing Classification System* e o *IEEE Taxonomy*. No Brasil temos o CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) que tem a tabela de áreas de conhecimento. Nesses exemplos citados, essas classificações além de estruturar e organizar a informação, também trazem relações entre áreas.

Escolha da Taxonomia e Seleção dos Artigos

É importante a escolha de uma determinada taxonomia, pois ela será a base de comparação com as recomendações que serão feitas na última etapa do método. Assim restringimos nossas comparações em apenas um universo e conseguimos resultados mais consistentes.

Outro ponto importante nesta etapa é a seleção de artigos, pois é dela que extrairemos as informações necessárias para criar o grafo na próxima etapa. As principais informações a serem extraídas são os autores do artigo e as áreas de atuação do trabalho, entretanto informações como filiação dos autores, ano de publicação, local de publicação etc, também podem ser extraídas e acrescentar mais informações à análise.

Criação do Grafo de Relacionamentos entre Áreas

O primeiro passo desta etapa é a criação do grafo de relacionamento de coautoria. A partir do corpus definido na etapa anterior, será criado um grafo onde os nós são os autores, as arestas são os artigos que escreveram juntos e o peso da aresta é a frequência com que esta parceria se repetiu (a quantidade de artigos que os autores possuem em comum).

A partir do sociograma, será traçado um grafo de correlação de áreas.

Um grafo G é um par ordenado (V, A) formado por um conjunto de |V| vértices, dado por V = {v1, v2, ..., v|V|}, e um conjunto de |A| arestas, dado por A = {a1, a2, ..., a|A|}, onde, no nosso caso, os vértices são as áreas de atuação de todos os autores e também áreas de todas as publicações. As arestas são formadas a partir de uma ligação entre coautores, ou seja, as áreas de atuação entre dois ou mais autores de uma mesma publicação são interligadas e as áreas de atuação das publicações também são interligadas. Para isto, é executado os seguintes passos:

i. Relacionar as áreas de atuação entre pesquisadores em coautoria

ii. Relacionar as áreas de atuação de cada pesquisador;

iii. Formar o grafo com todos esses relacionamentos;

Na Figura 3, temos dois pesquisadores e suas publicações. O pesquisador 1, tem as publicações Pub1 e Pub2. Já o pesquisador 2, tem as publicações definidas como Pub2 e Pub3. Além disso, as publicações têm áreas de atuação atreladas a elas. A Pub1 tem as áreas de atuação A1 e A2, Pub2 tem as áreas A3 e A4 e Pub3 tem relacionado a área A5. Como podemos perceber os pesquisadores 1 e 2 tem uma publicação em comum (Pub2), por isso eles têm uma conexão entre eles e as áreas A3 e A4 (0) aparecem já que são as áreas de Pub2. A imagem 2 ilustra os pesquisadores com suas publicações e as publicações com suas áreas de atuação.

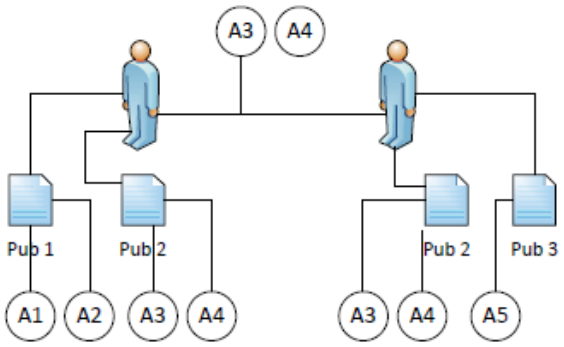

Figura 3 - Exemplo de ligação entre autores e suas publicações.

A Figura 4 mostra que o pesquisador 1 tem relação com as áreas A1, A2, A3 e A4 e o pesquisador 2 tem relação com A3, A4 e A5. Transformamos as publicações em áreas de atuação, visto que nosso grafo é de relacionamento entre áreas.

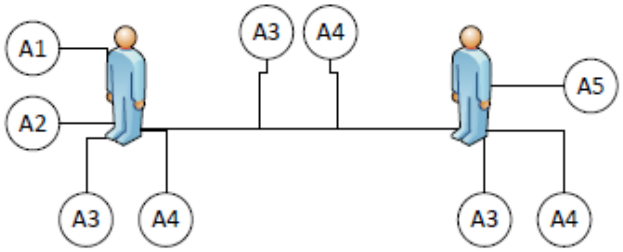

Figura 4 - Transformação das publicações em somente áreas de atuação

A partir dessa relação entre pesquisadores e áreas, podemos criar nosso grafo de relacionamento entre áreas. Temos nossos vértices V= {A1, A2, A3, A4 e A5} e o conjunto de arestas formadas pelo relacionamento entre áreas da mesma publicação e áreas que se relacionam pela coautoria entre pesquisadores, Aa= ({A1, A2}, {A1, A3}, {A1, A4}, {A2, A3}, {A3, A4}) são as relações das publicações do pesquisador 1. Ab= ({A3, A4}, {A3, A5}, {A4, A5}) são as relações entre as áreas do pesquisador 2. E Ac= ({A3, A4}) é a relação entre a publicação em comum entre os pesquisadores. Com isso unimos Aa, Ab e Ac e temos o conjunto total de arestas do grafo A = ({A1, A2}, {A1, A3}, {A1, A4}, {A2, A3}, {A3, A4}, ({A3, A4}, {A3, A5}, {A4, A5}, {A3, A4}). Observamos nesse exemplo que a relação entre as áreas A3 e A4 aparece 3 vezes. Essa propriedade do relacionamento em grafo será utilizada na próxima seção.

Seguindo o algoritmo e baseado no esquema da Figura 4, uniremos as áreas de atuação relacionadas a coautoria entre autores, Figura 5, (i), as áreas relacionadas aos

pesquisadores, Figura 5, (ii). Por fim, unimos todas as relações em um grafo único, Figura 5, (iii).

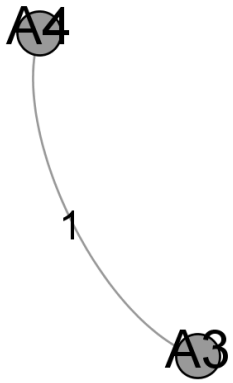

Figura 5 - Passo (i)

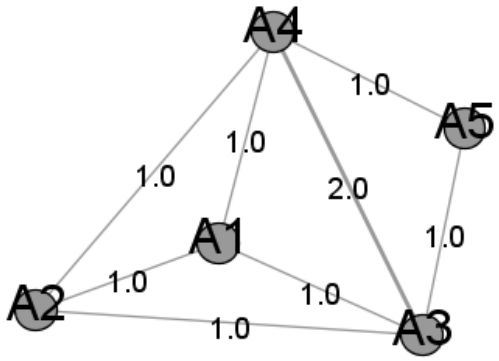

Figura 5 - Passo (ii)

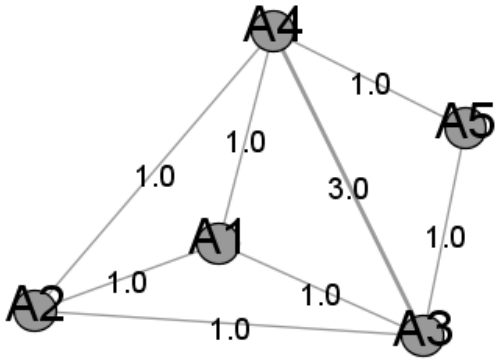

Figura 5 - Passo (iii)

Seleção das áreas

Nesta etapa, encontramos as conexões mais fortes entre as áreas de conhecimento através do peso das arestas. Por exemplo, conforme podemos ver na figura 5 (iii), a ligação entre os vértices A3 e A4 tem um peso 3, enquanto as outras em o peso 1. O peso é a quantidade da frequência que os vértices se relacionam. Com isso, podemos verificar que as áreas A3 e A4 tem uma relação de proximidade muito mais forte nesse grafo, do que as outras áreas.

Utilizamos um filtro baseado no peso das arestas (FiltroRelacÁreas) para estabelecermos os relacionamentos mais fortes entre as áreas.

$$Filtro_{RelacÁreas} = \frac{Pmax(|A|)}{2}$$

Equação 1 – Filtro de Peso das arestas

Sendo Pmax(|A|) o valor de maior peso encontrado no grafo. Ou seja, nosso filtro é baseado na metade do maior peso entre as arestas do grafo. Nesta etapa, selecionamos todas as áreas (vértices do grafo) cujas arestas são selecionadas através do $Filtro_{RelacÁreas}$.

Agrupamento entre áreas

O próximo passo é o agrupamento de áreas em subconjuntos.. Essa etapa é aplicada para a detecção de grupos de vértices que compartilham propriedades em comum e/ou desempenham funções semelhantes dentro do grafo com um grande volume de áreas. Caso o número de áreas resultantes seja pequeno após o filtro (etapa anterior), esta etapa poderá ser descartada. Se após o filtro de peso da aresta for realizado, restar uma quantidade abaixo de 10% de arestas do grafo original, então a etapa de agrupamento não deve ser realizada. Pois possivelmente as relações do subgrafo já tem alta densidade de conexões entre seus vértices. Por exemplo: o grafo de áreas tem 1000 arestas, após o filtro restam 98 arestas entre as áreas, nesse caso o agrupamento não é necessário.

Recomendação de Atualizações na Taxonomia

Para a recomendação de modificações para a taxonomia original, tentamos encontrar as arestas que tem os maiores pesos e analisamos as áreas desse relacionamento.

Primeiramente vamos definir conceitos importantes para o entendimento desta seção. Durante nossas pesquisas e taxonomias analisadas (ACM, IEEE e CNPq), geralmente uma taxonomia, é dividida em 4 conjuntos de conceitos (Grande Área, Área, Subárea e Especialidade) que se relacionam hierarquicamente indo dos mais abrangentes (Grande área) até os mais específicos (Especialidades). Esses conjuntos não são únicos e fixos, o aumento ou subtração desses conjuntos pode ser aceito se produzir resultados satisfatórios nos arranjos de classificação. O importante é manter a classificação hierárquica e o relacionamento de áreas mais abrangentes com áreas mais específicas.

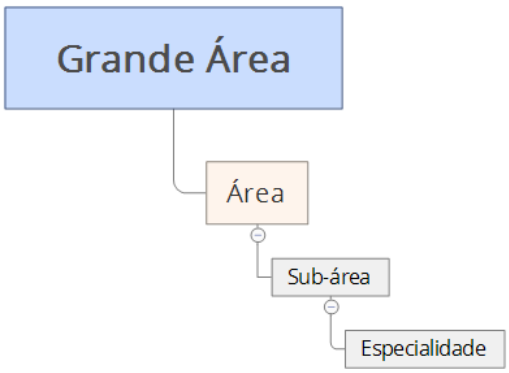

Figura 6 – Hierarquia genérica de uma taxonomia baseada nas bases estudadas (IEEE, ACM, CNPq)

Sendo assim, os passos para a recomendação de novas áreas são:

1. Identificação dos relacionamentos fortes e fracos. Os relacionamentos fortes são aqueles que tem o peso da aresta acima do valor de filtro (Equação 1). Os relacionamentos que têm o peso da aresta menor que os filtros são considerados relacionamentos fracos.

2. Criação da tabela de grau de mudança. Esta é uma tabela de que mapeará a mudança de cada área com as demais. Uma relação direta é quando as áreas fazem parte da mesma zona hierárquica. Por exemplo, áreas que são especialidades e fazem parte da mesma sub-área.

   Para preencher esta tabela, utilizaremos como parâmetros:

   I. Áreas com uma relação direta entre elas e não precisam sofrer uma modificação → grau de mudança = 0.

   II. Se a relação não for direta, mas ainda fazem parte da mesma Área→ grau de mudança =1

   III. Em um relacionamento entre áreas que não fazem parte da mesma Área, mas tem uma relação porque fazem parte da mesma grande área→ atribuímos o valor 2 para o grau de mudança.

   IV. Se uma existir uma relação entre duas áreas que não se enquadram em nenhum dos casos descritos anteriormente, então elas não têm nem uma grande área em comum → recebem o grau de mudança 3.

3. Com a tabela de grau de mudança preenchida, unimos a tabela de peso da aresta com a tabela de grau de mudança.

   - Ordenamos de forma decrescente os pesos das arestas. Com isso podemos ver os maiores pesos, ou seja, os relacionamentos mais fortes.
   - Nesta lista, buscamos dentre os relacionamentos mais fortes, os que tem o grau de mudança 3.
   - Verificamos em qual grupo (Grande áreas, Área, subárea ou especialidade) as áreas pertencem:

- Se elas são de grupos diferentes, a área que é mais específica fica conectada num nível abaixo àquela mais abrangente;
- Se são do mesmo grupo, o elaborador deve definir qual área é mais abrangente e definir a hierarquia.

Cabe ressaltar que toda taxonomia é um processo de representação e classificatório da informação e como todo processo desta natureza é um produto de uma construção que representa o estado e visão do conhecimento de seus elaboradores.

4. Resultados

Esta seção é apresentada no intuito de demonstrar nosso método com dados reais e utilizar uma taxonomia também real. Nossa escolha foi pela taxonomia do IEEE, porque além de ser referência para as áreas de engenharias e computação, sua última versão publicada é de 2014. São motivações mais do que suficientes para escolhermos essa taxonomia e propor uma atualização da mesma.

Os dados escolhidos para aplicar nossa metodologia foram da base da IEEE (Institute of Electrical and Electronics Engineers), analisamos a conferência ICDM (International Conference in Data Mining) e a IJCNN (International Joint Conference on Neural Networks).

Utilizamos o IEEE Xplore Search Gateway como ferramenta de exportação das informações. Essa API faz solicitações via HTTP à biblioteca do IEEE e retorna os valores em formato XML

Um exemplo de uma consulta[1]. Podemos traduzi-la da seguinte forma: retorne as publicações com ano de publicação acima de 2009 que contenham a palavra ‘java’ nos seus metadados e com ISSN igual a ‘1077-2626’.

Nosso foco é na análise e os relacionamentos das áreas de atuação dos pesquisadores. Para isso definimos que as áreas de atuação serão as palavras-chave definidas em cada publicação. O IEEE define 4 diferentes tipos de palavras-chave Figura 7:

i. Author Keywords;
ii. IEEE Terms;
iii. INSPEC: Controlled Indexing
iv. INSPEC: Non Controlled

[1] http://ieeexplore.ieee.org/gateway/ipsSearch.jsp?querytext=java&pys=2010&issn=1077-2626

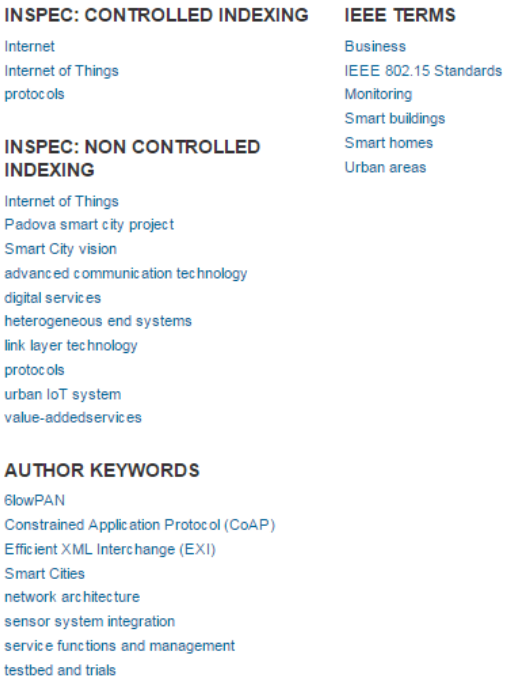

Figura 7 - Tipo de palavras-chave IEEE

Em (i) são as palavras cadastradas pelo(s) autor(es) na publicação, sendo de livre preenchimento e nenhuma padronização. Em (ii, iii e iv) os termos são gerados automaticamente baseados na análise de conteúdo do trabalho e utilizam uma organização padronizada. (ii) utiliza o IEEE Thesaurus, já (iii e iv) utilizam Inspec Thesaurus[2]. Definimos as palavras-chave geradas por (ii) como as áreas de atuação da publicação porque queremos analisar os termos que compõem a taxonomia do IEEE e assim atualiza-la.

Em todos nossos resultados utilizamos uma mesma base de dados para armazenar as informações. Na Figura 8 podemos ver a modelagem criada para receber os dados vindos da API do IEEE.

[2] http://www.theiet.org/resources/inspec/about/records/ithesaurus.cfm

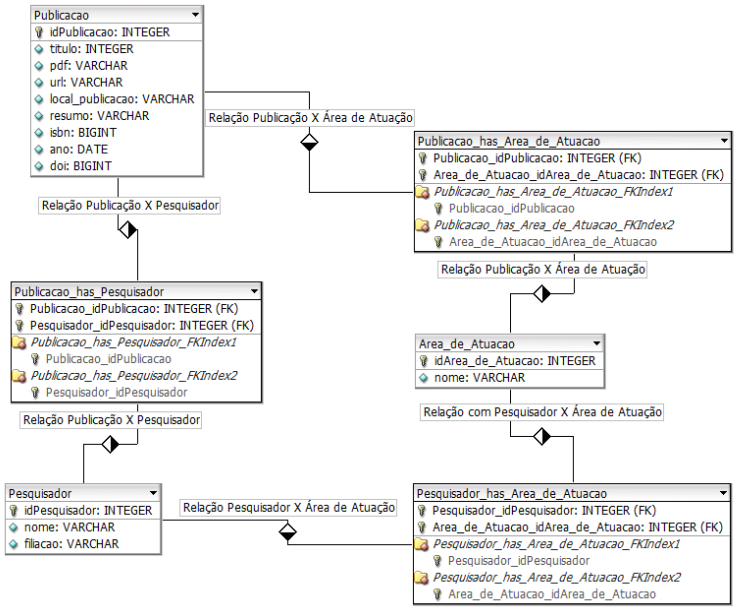

Figura 8 - Modelagem do banco de dados

A seguir serão apresentados os resultados que obtivemos onde aplicamos nossa metodologia em 2 diferentes grupos. Escolhemos o ICDM e IJCNN porque são duas conferências internacionalmente conhecidas.

ICDM

Nesta seção iremos apresentar os resultados da nossa metodologia na *International Conference on Data Mining* (ICDM). Nosso objetivo é aplicar o método em uma área mais específica para que possamos analisar uma parte mais específica da taxonomia. Escolhemos a ICDM porque ela é uma referência em pesquisas que envolvam a área de *data mining*, e tem estrato indicativo de qualidade da CAPES - A1.

Para extrair as informações e armazená-las no nosso banco de dados foi utilizado a consulta[3], que pode ser traduzida assim: *"retorne os 1000 trabalhos do ISSN[4]: 1550-4786, começando pelo ano de 2010"*. Com essa consulta foi possível extrair informações de 986 publicações. A Figura 9 a seguir mostra a distribuição das publicações por ano.

3 http://ieeexplore.ieee.org/gateway/ipsSearch.jsp?hc=1000&issn=1550-4786&pys=2010

4 International Standard Serial Number

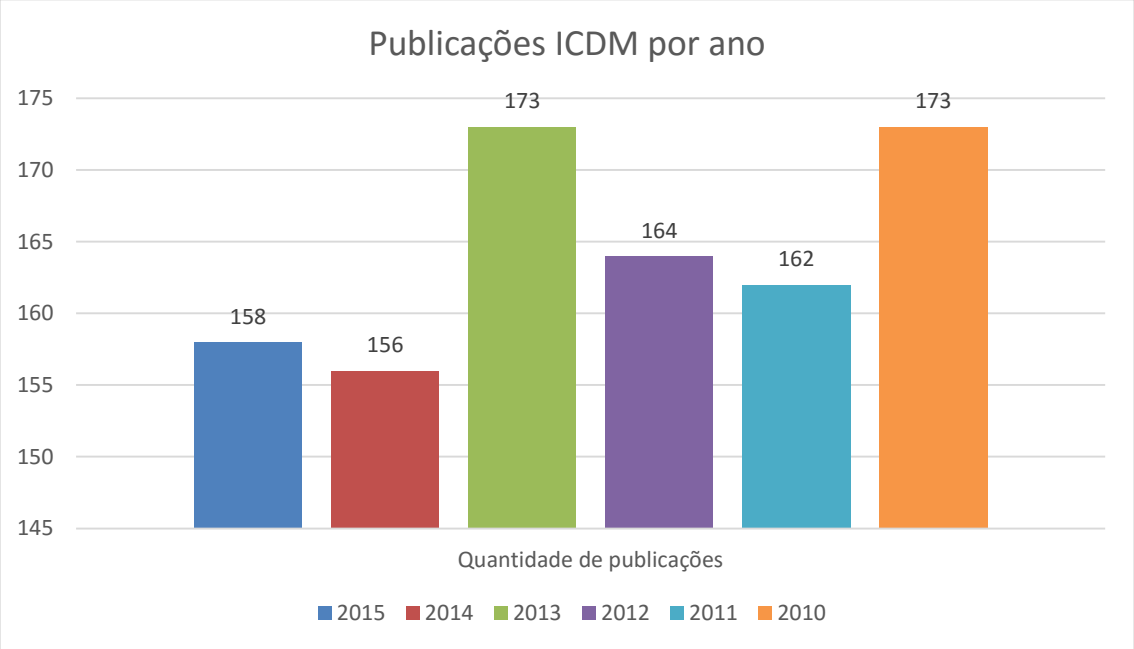

Figura 9 - Quantidade de publicações na ICDM

Após o armazenamento das informações de todas as publicações, criamos o grafo de relacionamento entre as áreas, levando em consideração a rede de coautoria. O grafo é composto por 650 nós e 26924 arestas. Uma visão geral do grafo completo pode ser vista na Figura 10.

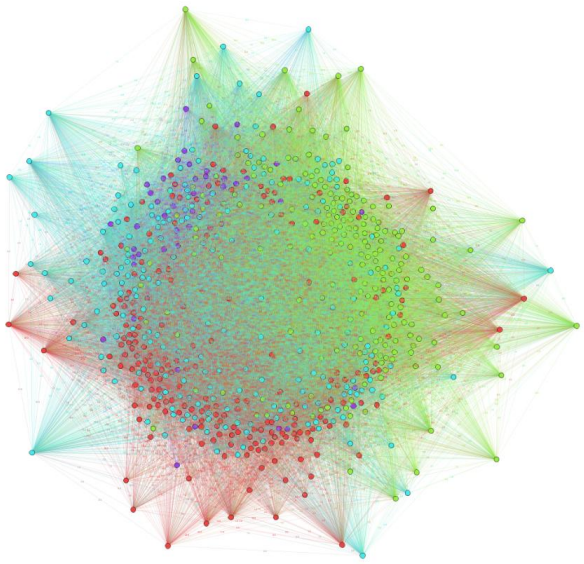

Figura 10 – Visão geral do Grafo ICDM

Ao analisar as relações entre os vértices obtemos os pesos das arestas e de acordo com nossa metodologia, podemos identificar relacionamentos fortes. No nosso grafo temos o maior peso o valor 441, o filtro de áreas fortes diz que devemos dividir o maior peso e filtrar apenas os relacionamentos acima desse valor, ou seja, acima de 220. A Figura 11 mostra as arestas consideradas mais fortes e seus respectivos nós.

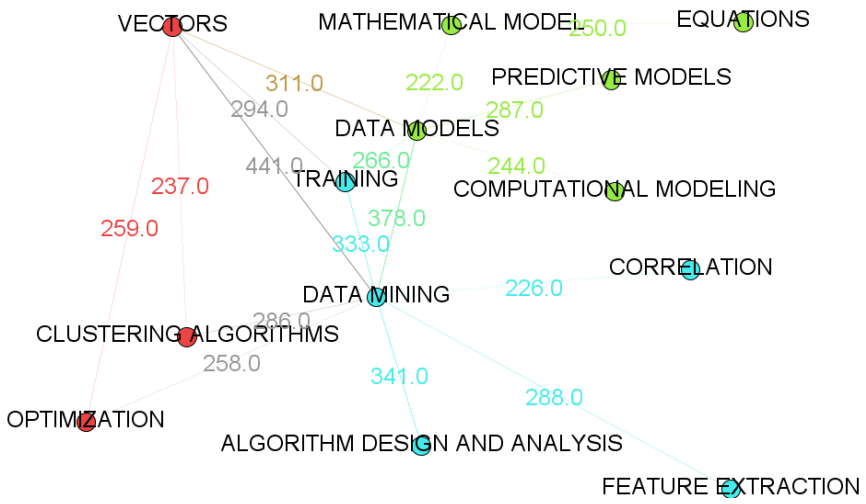

Figura 11 - Grafo ICDM filtrado

Como era de se esperar a área com maiores números de relacionamentos é data mining. A Tabela 1 mostra na integra os relacionamentos considerados 'fortes' após a realização do filtro e o grau de mudança atribuído de acordo com o posicionamento de cada área dentro da taxonomia, conforme explicado em 0.

Tabela 1 - Peso e grau de mudança dos relacionamentos fortes (ICDM)

Aresta (origem)	Aresta (destino)	Peso	Grau de mudança
DATA MINING	VECTORS	441	3
DATA MINING	DATA MODELS	378	3
ALGORITHM DESIGN AND ANALYSIS	DATA MINING	341	3
DATA MINING	TRAINING	333	3
DATA MODELS	VECTORS	311	3
TRAINING	VECTORS	294	3
DATA MINING	FEATURE EXTRACTION	288	2
DATA MODELS	PREDICTIVE MODELS	287	3
CLUSTERING ALGORITHMS	DATA MINING	286	3
DATA MODELS	TRAINING	266	3
OPTIMIZATION	VECTORS	259	2

DATA MINING	OPTIMIZATION	258	3
EQUATIONS	MATHEMATICAL MODEL	250	1
COMPUTATIONAL MODELING	DATA MODELS	244	1
CLUSTERING ALGORITHMS	VECTORS	237	2
CORRELATION	DATA MINING	226	3
DATA MODELS	MATHEMATICAL MODEL	222	3

Antes de propor as recomendações entre as áreas, iremos analisar a estrutura atual da área central desse experimento que é a área de *data mining.* A Figura 12 traz a estrutura atual onde se encontra a área de data mining na taxonomia do IEEE. Vemos que ela está relacionada à grande área chamada '*Computers and information processing*' e é uma subárea de *'Pattern recognition*'. E relacionadas diretamente com ela temos as seguintes especialidades: '*Association rules*', '*Data privacy'*, '*Text Analysis*', '*Text mining*' e *'Web mining*'. Dentre os relacionamentos determinados como fortes que encontramos, não observamos nenhum relacionamento entre essas áreas que já fazem parte de *data mining* na taxonomia. Isso nos leva a crer que os relacionamentos podem estar desatualizados.

Entre as áreas analisadas temos a reunião de grandes áreas e suas subáreas:

'***computional and artificial intelligence****': {predictive models},*

*'**computers and information processing**': {data mining; feature extraction},*

*'**superconductivity**': {data models},*

*'**education'**: {training} e*

*'**mathematics'**: {vectors; algorithms; clustering algorithms; algorithm design and analysis; equations; mathematical model; optimization; correlation}.*

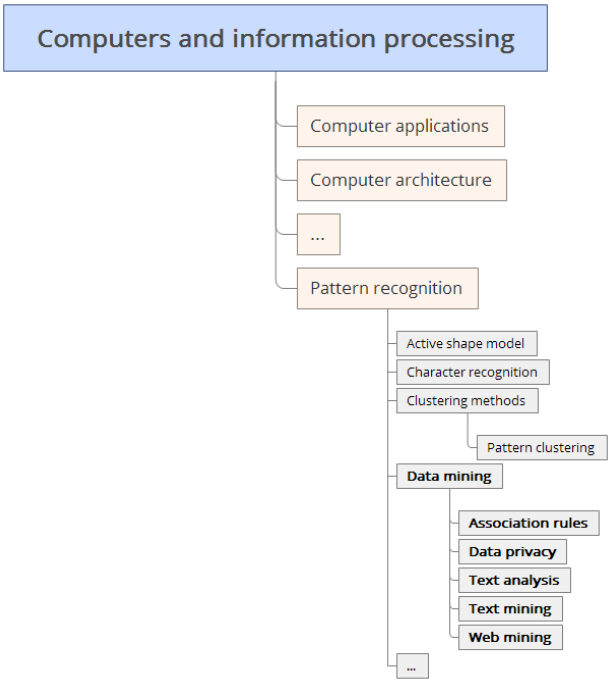

Figura 12 - Taxonomia original do IEEE - Área de Data Mining

Neste ponto, nosso objetivo é trabalhar com a tabela de pesos e grau de mudança, para que possamos recomendar novos relacionamentos. Como primeira análise vamos explorar os relacionamentos referentes à área de data mining, Tabela 2.

Tabela 2 - Relacionamentos (Data Mining)

Aresta (origem)	Aresta (destino)	Peso	Grau de mudança
DATA MINING	VECTORS	441	3
DATA MINING	DATA MODELS	378	3
DATA MINING	ALGORITHM DESIGN AND ANALYSIS	341	3
DATA MINING	TRAINING	333	3
DATA MINING	CLUSTERING ALGORITHMS	286	3
DATA MINING	OPTIMIZATION	258	3
DATA MINING	CORRELATION	226	3
DATA MINING	FEATURE EXTRACTION	288	2

Ao analisarmos os relacionamentos, vimos que não existe nenhum grau de mudança abaixo de 2 e então, consideramos todos os relacionamentos como válidos. Na taxonomia do IEEE a área de ‘data mining’ tem nível de subárea, por isso decidimos colocar todos os relacionamentos como seus ‘filhos’. A Figura 13 demonstra graficamente a nova estrutura com os relacionamentos.

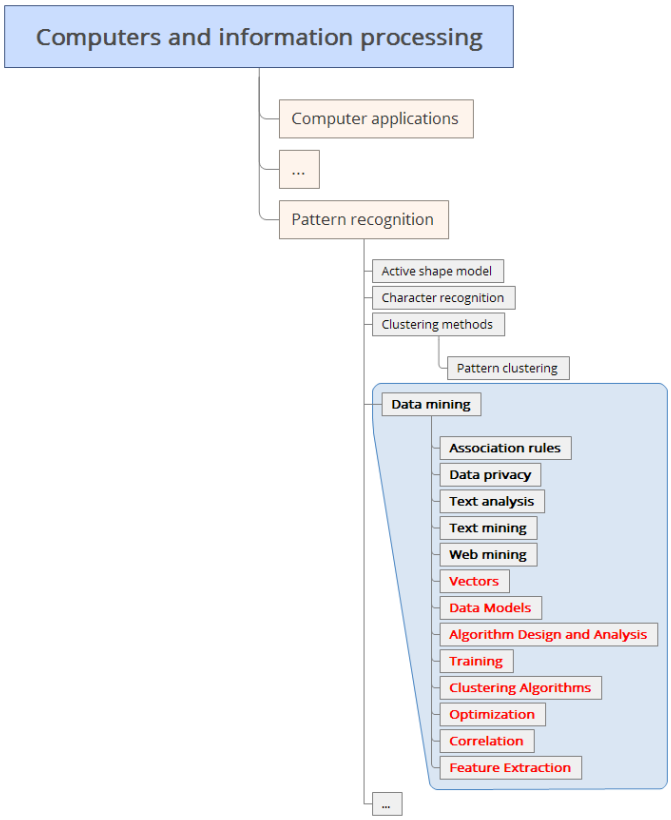

Figura 13 - Recomendação para 'Data Mining'

Além da área de *data mining*, iremos analisar as suas áreas correlatas, vimos uma forte relação entre a área '*data models'* com as áreas '*predictive models'* e '*mathematical model'*, conforme visto na Tabela 3.

Tabela 3- Relacionamentos (Data Models)

Aresta (origem)	Aresta (destino)	Peso	Grau de mudança
DATA MODELS	PREDICTIVE MODELS	287	3
DATA MODELS	MATHEMATICAL MODEL	222	3

A área de *'data models' está* ligada a grande área '*superconductivity'* e a área *'modeling',* consideramos as áreas '*mathematical model' e 'predictive models'* mais específicas e as introduzimos como uma especialidade de '*data models', ilustrado na* Figura *14.*

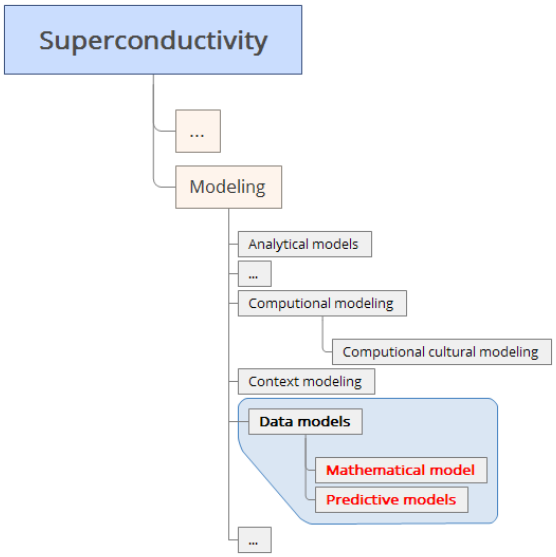

Figura 14 - Recomendação *Data Models*

Outra área que merece destaque é '*vectors'* ela foi junto com '*data mining'* o relacionamento mais evidente e forte da nossa análise, por isso demos destaque para ela. A Tabela 4 mostra esses relacionamentos.

Tabela 4 - Relacionamentos (Vectors)

Aresta (origem)	Aresta (destino)	Peso	Grau de mudança
DATA MINING	VECTORS	441	3
DATA MODELS	VECTORS	311	3
TRAINING	VECTORS	294	3
OPTIMIZATION	VECTORS	259	2
CLUSTERING ALGORITHMS	VECTORS	237	2

Temos as áreas '*data mining'* e '*data models'* que já foram mencionadas acima. Outras áreas como '*training'*, '*optimization'* e '*clustering algorithms'* também tem forte relação com '*vectors'*. '*Training'* é uma área e tem relação com a grande área '*Education'*. Acreditamos que o contexto de *training* utilizado pela IEEE na taxonomia não é o mesmo quando pensamos em *data mining.* Porém não iremos modificar a sua estrutura, vamos nos adequar os novos conceitos à realidade. '*Optimization'* e '*Clustering Algorithms'* são áreas relacionadas com a grande área '*Mathematics'*.

Decidimos utilizar a área *training* como referência para as demais áreas, pois dentro da hierarquia da taxonomia do IEEE ela é uma área. Relacionamos diretamente ao '*training'* a área de '*data mining'*, mais uma vez por considerar seu conceito mais abrangente. Em seguida conectamos as outras áreas como filhas de *data mining, repetindo o processo da* Figura *13,* só que nesse caso levando em consideração

somente as áreas relacionadas com '*vectors'. O exemplo da estrutura pode ser visto na* Figura *5*.

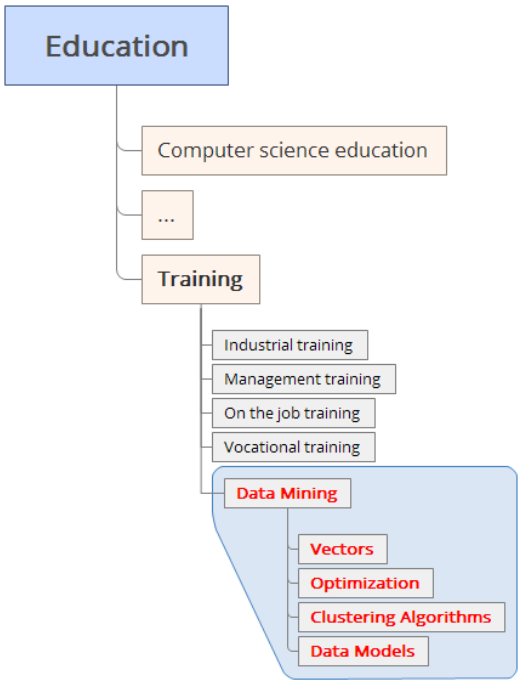

Figura 15 – Recomendação *Training*

Para finalizar decidimos mostrar um caso de falso forte. Falso forte é aquele relacionamento que pelo peso na aresta entre as áreas se mostra como um relacionamento forte. Entretanto ao analisarmos o grau de mudança do relacionamento vimos que eles fazem parte da mesma estrutura, mesmo nível na hierarquia e estão relacionadas com a mesma área. Por esses motivos têm grau de mudança igual a 1. Então apesar de serem considerados como um relacionamento forte, são falsos, pois seu relacionamento já existe na taxonomia. A Tabela 5 mostra um exemplo, '*Computional modeling'* e '*Data models'* tem a mesma grande área, '*superconductivity'* e também fazem parte da mesma área '*Modeling'*. É o mesmo caso para '*Equations'* e '*Mathematical model'* que fazem parte da mesma grande área '*Mathematics'*.

Tabela 5 - Relacionamentos falso forte

Aresta (origem)	Aresta (destino)	Peso	Grau de mudança
COMPUTATIONAL MODELING	DATA MODELS	244	1
EQUATIONS	MATHEMATICAL MODEL	250	1

IJCNN

Conforme utilizamos uma área de pesquisa mais específica na seção 0, nesta seção também vamos explorar uma nova área. Iremos aplicar nossa metodologia na

International Joint Conference on Neural Networks (IJCNN). Escolhemos essa conferencia porque ela era direcionada para uma área de atuação e tem estrato da CAPES – A2. Consideramos que os melhores trabalhos da área de redes neurais podem estar reunidos nesta conferência. E analisar os melhores trabalhos juntamente com os pesquisadores que publicaram para essa conferencia pode nos trazer relacionamentos interessantes para a área de redes neurais.

Para extrairmos as informações utilizamos a seguinte consulta[5], que se traduz da seguinte maneira: retorne as 1000 publicações do ano de 2015 a conferencia com o ISSN igual a 2161-4393. Após a consulta, armazenamos as informações e obtivemos um total de 560 publicações. A Figura 6 mostra uma visão geral do grafo de relacionamentos entre áreas dos pesquisadores em coautoria no IJCNN.

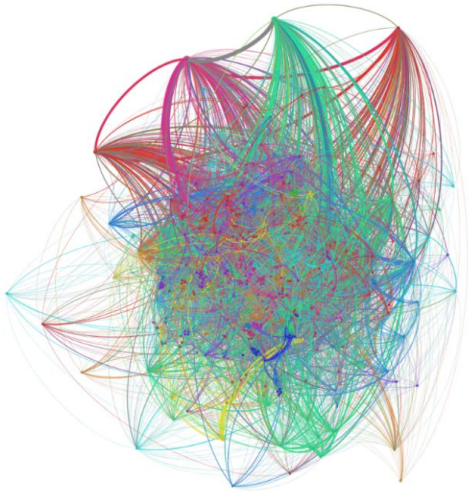

Figura 16 - Visão geral do Grafo IJCNN

Para analisarmos os relacionamentos mais importantes, seguindo nossa metodologia, aplicamos o filtro nas arestas. O maior peso encontrado entre as relações foi 59, nosso filtro foi estabelecido em 28. Ou seja, relacionamentos com peso acima de 29 foram considerados 'fortes'. A Figura 7 mostra o grafo filtrado com os relacionamentos fortes e a Tabela 6 mostra os relacionamentos ordenados pelo peso na aresta.

[5] http://ieeexplore.ieee.org/gateway/ipsSearch.jsp?hc=1000&issn=2161-4393&py=2015

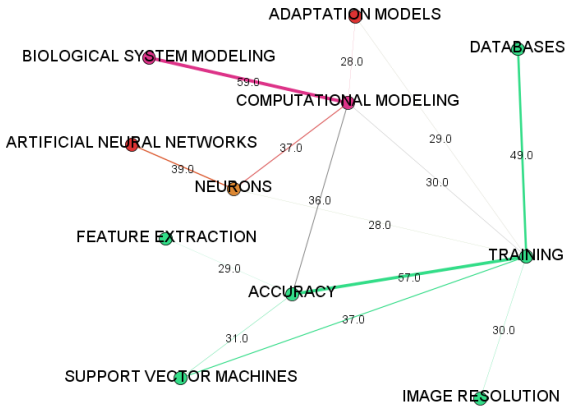

Figura 17 - Grafo IJCNN filtrado

Tabela 6 - Relação entre áreas mais fortes (IJCNN)

Aresta (Origem)	Aresta (Destino)	Peso
BIOLOGICAL SYSTEM MODELING	COMPUTATIONAL MODELING	59
ACCURACY	TRAINING	57
SOCIOLOGY	STATISTICS	52
DATABASES	TRAINING	49
ARTIFICIAL NEURAL NETWORKS	NEURONS	39
COMPUTATIONAL MODELING	NEURONS	37
SUPPORT VECTOR MACHINES	TRAINING	37
BRAIN MODELING	ELECTROENCEPHALOGRAPHY	36
ACCURACY	COMPUTATIONAL MODELING	36
ACCURACY	SUPPORT VECTOR MACHINES	31
COMPUTATIONAL MODELING	TRAINING	30
HEART	IRIS	30
IMAGE RESOLUTION	TRAINING	30
BIOINFORMATICS	GENOMICS	30
ADAPTATION MODELS	TRAINING	29
GLASS	IRIS	29
ACCURACY	FEATURE EXTRACTION	29

Na figura e tabela supracitada temos os relacionamentos mais fortes, o relacionamento com o maior peso é entre '*biological system modeling*' e '*computational modeling'*. Temos outros relacionamentos interessantes como '*accuracy'* e '*training'*, ou '*artificial neural networks' e 'neurons'*. Para criarmos a tabela com o grau de mudança para cada relacionamento precisamos analisar a estrutura hierárquica de cada

área. Dentre as áreas mais fortes temos as seguintes grandes áreas, em negrito, com suas respectivas subáreas:

Computational and artificial intelligence: *{artificial neural networks, support vector machines};*

Computers and information processing: *{feature extraction, image resolution};*

education*: {training};*

engineering in medicine and biology*: {bioinformatics, biological system modeling, genomics};*

instrumentation and measurement*: {electroencephalography};*

materials elements and compounds*: {glass};*

mathematics*: {accuracy, adaptation models, statistics};*

professional communication*: {databases};*

science – general*: {sociology};*

superconductivity*: {brain modeling , computational modeling}.*

*As áreas {****heart, iris*** e ***neurons****}* fazem parte apenas do tesauro do IEEE.

Com isso podemos acrescentar à nossa tabela a coluna com o grau de mudança.

Tabela 7 - Relacionamentos (IJCNN)

Aresta (Origem)	**Aresta (Destino)**	**Peso**	**Grau de mudança**
BIOLOGICAL SYSTEM MODELING	COMPUTATIONAL MODELING	59	3
ACCURACY	TRAINING	57	3
SOCIOLOGY	STATISTICS	52	3
DATABASES	TRAINING	49	3
ARTIFICIAL NEURAL NETWORKS	NEURONS	39	3
COMPUTATIONAL MODELING	NEURONS	37	3
SUPPORT VECTOR MACHINES	TRAINING	37	3
BRAIN MODELING	ELECTROENCEPHALOGRAPHY	36	3
ACCURACY	COMPUTATIONAL MODELING	36	3
ACCURACY	SUPPORT VECTOR MACHINES	31	3
COMPUTATIONAL MODELING	TRAINING	30	3
HEART	IRIS	30	3

IMAGE RESOLUTION	TRAINING	30	3
BIOINFORMATICS	GENOMICS	30	2
ADAPTATION MODELS	TRAINING	29	3
GLASS	IRIS	29	3
ACCURACY	FEATURE EXTRACTION	29	3

Em nossos relacionamentos encontramos apenas *Bioinformatics* e *Genomics* quem fazem parte da mesma grande área, mas não da mesma área, por esse motivo tem grau de mudança igual a 2. Já os outros ganham grau 3 por não ter nenhuma relação hierárquica na taxonomia do IEEE.

Antes de propor recomendações de relacionamentos de atualização para a taxonomia, iremos analisar como é atualmente definido a área de redes neurais na taxonomia do IEEE. A Figura 8 mostra a estrutura hierárquica da área. Temos que '*Neural Networks*' é uma área ligada à grande área *'Computional and artificial intelligence*' e tem 8 subáreas interligadas diretamente que são: *'Artificial neural networks', Biological neural networks', Cellular neural networks', Feedforward neural networks', 'Multi-layer neural network', 'Neural network hardware', 'Radial basis function networks'* e *'Recurrent neural networks'*. E as especialidades interligadas indiretamente: *'Hebbian theory', Self-organizing feture maps', 'Multilayer perceptrons', 'Hopfield neural networks'*. Entre as áreas consideradas fortes pelo nosso estudo, apenas a '*Artificial Neural Networks'* já faz parte da hierarquia de redes neurais na taxonomia do IEEE.

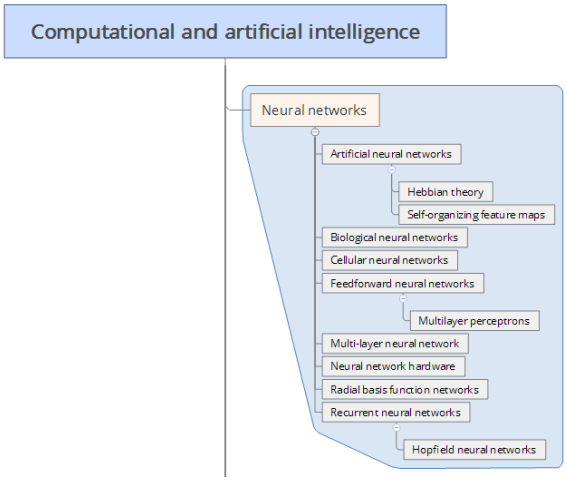

Figura 18 - Taxonomia do IEEE (Neural Networks)

Uma das áreas que mais chamou atenção foi '*Computional modeling'* pois aparece entre as áreas mais fortes. Ao analisarmos seus relacionamentos chegamos a Tabela 8, outra área que consideramos com grande relação foi '*brain modeling'*. Ainda na Tabela 8 temos conceitos como '*neurons*' e '*electroencephalography'* que a princípio

parecem ter mais uma relação médica, mas ao levarmos para o contexto de redes neurais fazem todo o sentido.

Tabela 8 – Relacionamentos Computational Modeling

Aresta (Origem)	Aresta (Destino)	Peso	Grau de mudança
BIOLOGICAL SYSTEM MODELING	COMPUTATIONAL MODELING	59	3
COMPUTATIONAL MODELING	NEURONS	37	3
ACCURACY	COMPUTATIONAL MODELING	36	3
BRAIN MODELING	ELECTROENCEPHALOGRAPHY	36	3
COMPUTATIONAL MODELING	TRAINING	30	3
ADAPTATION MODELS	COMPUTATIONAL MODELING	28	3

Consideramos a área '*Computional modeling'* a área mais abrangente e por isso, as áreas relacionadas foram posicionadas hierarquicamente como 'flhas' dela. A Figura 9 demostra a posição delas na taxonomia, além de posicionar a área '*electroencephalography'* filha de '*Brain Modeling'* por considerarmos essa última mais abrangente.

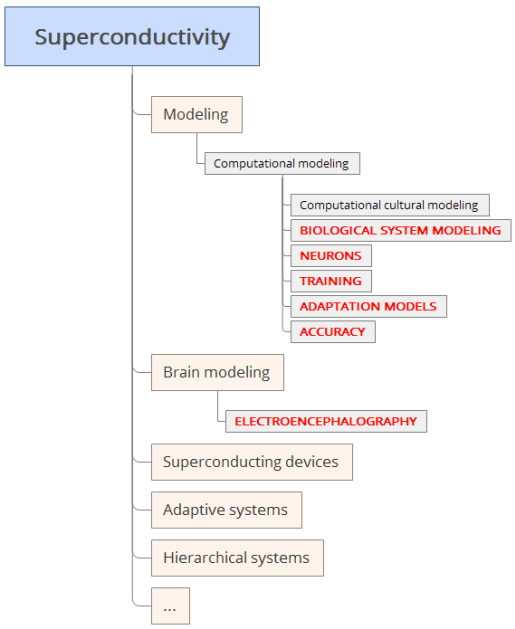

Figura 19 - Taxonomia proposta *Computional modeling* e *Brain Modeling*

Outra área bem evidente foi '*Training'* que na taxonomia do IEEE está intimamente ligada somente à área da Educação. Porém entendemos que ela é parte importante dos estudos de redes neurais. As áreas que fazem relação com *training* nesse

contexto de *neural networks* são *accuracy, databases, SVM, computional modeling, image resolution,* mostrado na *Tabela 9*.

Tabela 9 - Relacionamentos (Training)

Aresta (Origem)	Aresta (Destino)	Peso	Grau de mudança
ACCURACY	TRAINING	57	3
DATABASES	TRAINING	49	3
SUPPORT VECTOR MACHINES	TRAINING	37	3
COMPUTATIONAL MODELING	TRAINING	30	3
IMAGE RESOLUTION	TRAINING	30	3
ADAPTATION MODELS	TRAINING	29	3
NEURONS	TRAINING	28	3

Conforme já mencionado, a área de *training* faz parte, dentro da taxonomia atual, da grande área *Education,* decidimos criar a subárea *neural networks* diretamente ligada à *training* e conectar os relacionamentos à subárea. O resultado da taxonomia com os novos relacionamentos é mostrado na **Figura** 20.

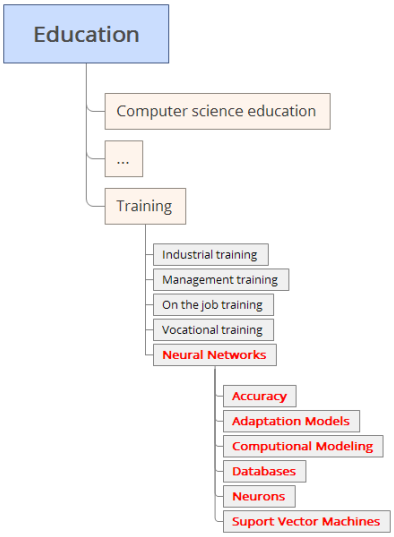

Figura 20 – Taxonomia proposta - Training

Em nossas análises encontramos um relacionamento que nos chamou atenção pela diferença entre as áreas. Um relacionamento forte entre '*sociology'* e '*statistics'* que aparentemente não tem uma relação visível, porém ao analisar o contexto

de redes neurais elas aparecem com grande relação. A Tabela 10 e a Figura 21 mostra, respectivamente, o relacionamento e a taxonomia proposta para *statistics* e *sociology.*

Tabela 10 – Relacionamento Sociology e Statistics

Aresta (Origem)	Aresta (Destino)	Peso	Grau de mudança
SOCIOLOGY	STATISTICS	52	3

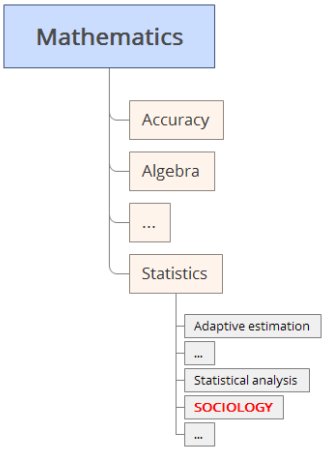

Figura 21 – Taxonomia proposta *Statistics* e *Sociology*

Para finalizar nossa análise e proposta da área de *neural networks* decidimos aplicar as propostas apresentadas anteriormente conectadas diretamente com o *neural networks.* O resultado é apresentado na Figura *22,* onde a área de *Neural Networks*, tem como novas subáreas *computional modeling* e *training*.

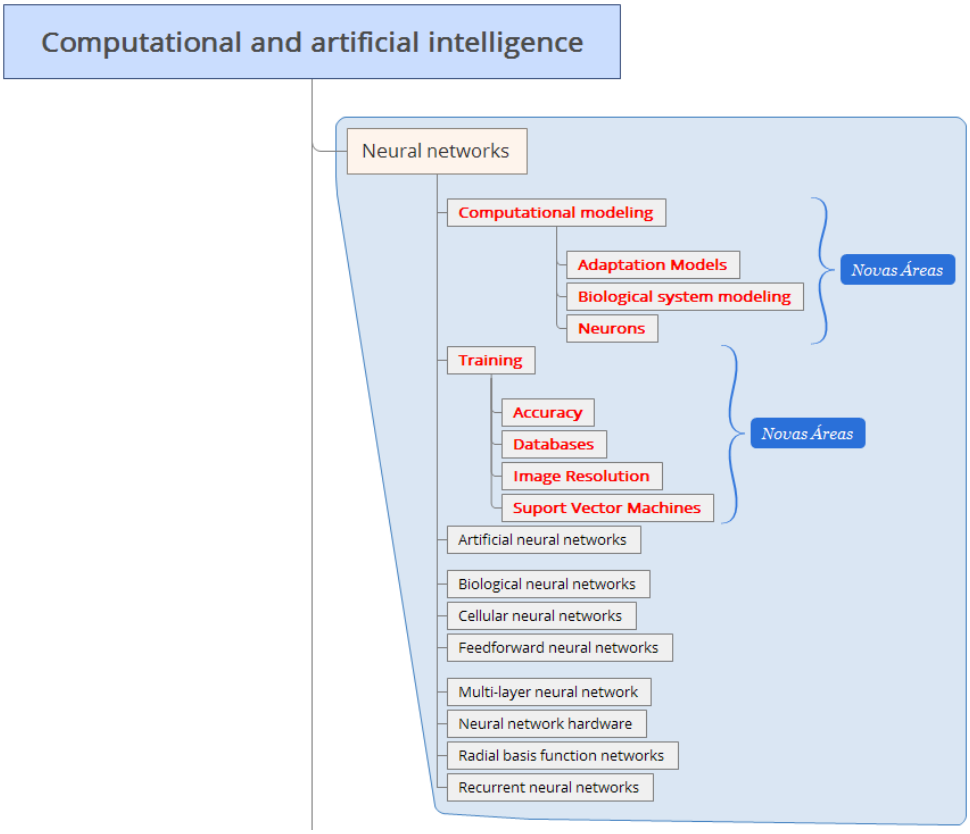

Figura 22 – Taxonomia proposta – Neural Networks

5. Avaliação

A avaliação foi aplicada no dia 16 de setembro de 2016 com o grupo do Laboratório CORES (Laboratório de Computação Social e Análise de Redes Sociais) e professores da UFRJ/PPGI, conseguimos a presença de 11 pessoas que avaliaram os resultados do trabalho feio nas duas bases mais específicas: *Data Mining* e *Neural Networks.*

Perfil dos participantes

A maioria dos avaliadores eram estudantes de graduação e pós-graduação, a Tabela 11 e a Figura 23 mostram a quantidade por tipo de formação dos avaliadores. Dentre eles, cerca de 55% têm experiência mediana ou superior no assunto *Data Mining* e 55% têm pouca ou mais experiência em *Neural Networks.* A Tabela 12 mostra um panorama da experiência dos avaliadores. A maioria dos avaliadores tem igual ou mais de 2 anos de experiência no assunto de Data Mining e mais de 1 ano de experiência em Neural Network. Com isso os avaliadores foram considerados pessoas capacitadas para analisar os resultados da metodologia. A Tabela 13 mostra a experiência de cada avaliador nos temas analisados.

Tabela 11 – Perfil dos avaliadores

Perfil	Qtd.	%
Graduando	4	36,36%
Mestrando	5	45,45%
Doutorando	2	18,18%
Total	11	100%

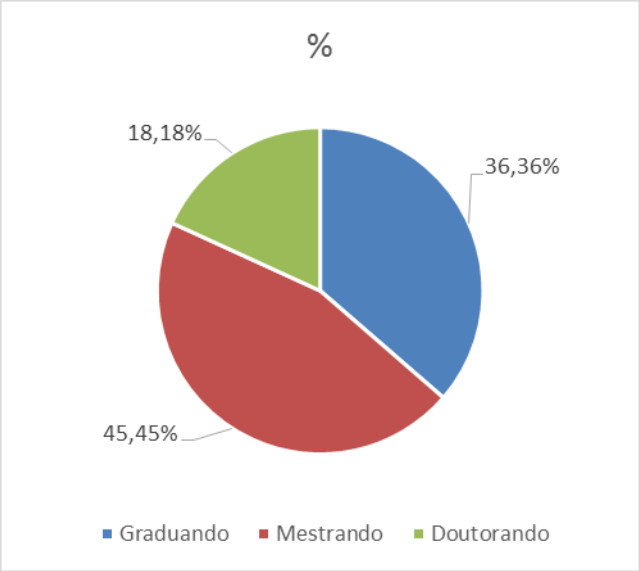

Figura 23 – Perfil dos avaliadores

Tabela 12 – Panorama das experiências dos avaliadores

Data Mining		Neural Networks	
0%	Nenhuma	45%	Nenhuma
45%	Pouca	36%	Pouca
27%	Média	0%	Média
18%	Alta	9%	Alta
9%	Expert	9%	Expert

Tabela 13 – Experiências dos avaliadores

Avaliador	Exp. (DM)	Exp. (NN)
1	Pouca	Nenhuma
2	Média	Pouca
3	Pouca	Nenhuma

4	Pouca	Nenhuma
5	Expert	Expert
6	Média	Pouca
7	Pouca	Nenhuma
8	Alta	Alta
9	Alta	Pouca
10	Média	Pouca
11	Pouca	Nenhuma

Análise Quantitativa e Qualitativa

Queremos analisar o *feedback* realizado pelos avaliadores, para isso foram realizadas perguntas sobre a estrutura, relacionamento e novas áreas. Os avaliadores podiam analisar a taxonomia atual e comparar com a taxonomia nova. Também foi realizado perguntas objetivas, com o intuito de saber:

i. Se as novas áreas estavam relacionadas com o domínio;
ii. Se as áreas estavam posicionadas hierarquicamente corretamente;
iii. E, se o avaliador desejaria relacionar a área analisada em outro domínio.

A seguir mostramos os resultados divididos para cada domínio analisado.

Data Mining

Entre os especialistas avaliados, 93% acham que as novas áreas relacionadas ao domínio estão corretas. E 81% acham que elas estão posicionadas corretamente em relação ao domínio de *Data Mining*. A

Tabela 14 traz o resumo da avaliação de cada avaliador.

Tabela 14 – Tabela de avaliação (*Data Mining*)

Avaliador	Domínio	Posição	Exp. (DM)
1	100%	100%	Pouca
2	100%	100%	Média
3	100%	100%	Pouca
4	100%	85%	Pouca
5	92%	62%	Expert
6	93%	67%	Média
7	80%	53%	Pouca

8	73%	73%	Alta
9	93%	80%	Alta
10	93%	80%	Média
11	93%	93%	Pouca
Média	**93%**	**81%**	

Em relação ao domínio, vemos que todos os avaliadores se mantiveram acima de 70%. Já em relação a posição notamos que os avaliadores 5, 6 e 7 foram os que tiveram uma avaliação abaixo dos 70%, dentre eles temos o único avaliador que se considerou um expert no assunto. Apesar disso, consideramos que tanto em relação ao domínio quanto em relação a posição na hierarquia, as novas áreas se incorporaram a taxonomia.

Em relação a outras áreas que o avaliador poderia propor ou explicar porque a área não fazia parte do domínio ou posição destacamos a área *Data Models*, alguns diziam que ela era muito abrangente ou que ela estaria melhor posicionada se estivesse conectada à Banco de Dados.

Neural Networks

Entre os especialistas avaliados, 86% acham que as novas áreas relacionadas ao domínio estão corretas. E 68% acham que elas estão posicionadas corretamente em relação ao domínio de *Neural Networks*. A Tabela 15 traz o resumo da avaliação de cada avaliador.

Tabela 15 - Tabela de avaliação (*Neural Network*)

Avaliador	Domínio	Posição	Exp. (NN)
1	93%	93%	Nenhuma
2	100%	100%	Pouca
3	100%	100%	Nenhuma
4	?	?	Nenhuma
5	100%	33%	Expert
6	87%	13%	Pouca
7	33%	33%	Nenhuma
8	80%	71%	Alta
9	80%	73%	Pouca

10	87%	73%	Pouca
11	100%	86%	Nenhuma
Média	**86%**	**68%**	

Na tabela acima vemos que o avaliador 4 não soube responder nenhuma pergunta ou preferiu se omitir. Esse posicionamento talvez seja pela falta de experiência no assunto. Ao analisarmos as respostas em relação ao domínio, apenas o avaliador 7 obteve média abaixo de 80%. Ao verificarmos as avaliações sobre o posicionamento das novas áreas, mais uma vez os avaliadores 5, 6 e 7 aparecem com a menor média de aprovação. O avaliador expert está entre eles e considerou 33% das novas áreas no posicionamento correto.

Entre os especialistas mais experientes os termos *Computional Modeling* e *Databases* são termos muito abrangentes e não deveriam estar posicionados no local que estavam na hierarquia.

A área de Neural Network obteve uma avaliação menor em relação à Data Mining nas avaliações de domínio e Posição. Avaliamos que isso pode ser pelo fato de que os avaliadores são mais experientes em Data Mining. Outros problemas que foram enfrentados durante a aplicação da avaliação são mostrados na seção a seguir.

Problemas

Nas revisões do formulário de avaliação foram encontrados problemas de entendimento da avaliação e uma melhoria na disposição das imagens foi proposta. Também foi remodelada a opção de escolher se a área faz parte do domínio e se ela estava posicionada corretamente. Após a primeira avaliação, cada área terá uma opção individual, podendo assim melhorar nossa avaliação do resultado da avaliação. Um problema mais geral identificado foi a complexidade de avaliar uma taxonomia, nenhum dos entrevistados é um especialista em organização da informação. Apesar disso ótimos resultados foram obtidos.

6. Considerações finais

Baseado em todo crescimento da produção acadêmica nos últimos anos e a dificuldade em organizar toda essa informação para uma melhor recuperação da informação ou até uma atualização das estruturas das áreas de conhecimento, esse estudo traz um método que apoiado no estudo da análise de redes sociais e organização do conhecimento pretende auxiliar na atualização da taxonomia das áreas da ciência.

Esse método é composto por 5 grandes etapas: escolha da taxonomia e seleção dos artigos, criação de um grafo de relacionamento entre áreas, seleção das áreas, recomendação de atualizações na taxonomia e avaliação.

Em todas as bases criadas conseguimos recomendar atualizações para a taxonomia que, posteriormente, foram avaliadas por especialistas.

A avaliação foi o processo que pôde consolidar todo o trabalho realizado e ratificar as etapas do método. De forma geral os resultados foram bem avaliados pelos especialistas, em alguns casos obtivemos 93% de média de acerto nas recomendações de novas áreas. Porém sabemos que podemos melhorar alguns aspectos que também foram levantados na etapa de avaliação.

Diante do exposto, acreditamos que o método tem muito a acrescentar na área da organização da informação no que diz respeito à atualização de uma taxonomia, principalmente por poder ser utilizado em qualquer domínio (Computação, Medicina, Engenharias etc).

Contribuições para a área

Como vimos anteriormente, as áreas de atuação precisam de atualização e o nosso trabalho traz mais uma vez essa discussão e propõe um novo método para auxiliar nessa atualização, baseando-se nas análises de redes sociais dos pesquisadores. O método apresentado é genérico e independente de tecnologia, por isso, pode ser replicado em qualquer área de atuação que deseje analisar e propor uma atualização para sua organização do conhecimento.

Limitações do estudo

O estudo encontrou algumas limitações e a melhor forma de apresenta-las foi em forma de perguntas e respostas.

Pergunta: Não ficou claro como são resolvidos os problemas de ambiguidade de nomes de autores?

Resposta: Como nossos testes foram realizados usando somente a biblioteca do IEEE, esperamos que a própria biblioteca trate e evite esse tipo de problema. Mas será incluído esse item como um ponto para trabalho futuro.

Pergunta: E se o grupo de pesquisadores que estou analisando não trabalha em coautoria ou tem poucos pesquisadores que trabalham em conjunto?

Resposta: Para esse tipo de situação, podemos utilizar outros sociogramas para gerar o nosso grafo de relacionamentos. Temos como opções mais interessantes que podem ser testadas a rede de citação, cocitação e acoplamento bibliográfico.

Pergunta: E se as publicações que estou extraindo as informações não tem palavras-chave?

Resposta: Existem diversas formas de extrair os temas mais importantes de uma publicação, recomendamos o *Topic Labeling* (NOLASCO, D.; OLIVEIRA, J., 2016).

Os trabalhos realizados por (D. BLEI, L. CARIN AND D. DUNSON, 2010; J.H. LAU ET AL, 2010 E 2011; D. MAGATTI, 2009) são boas referências no assunto.

Outra limitação encontrada foi o tempo, não conseguimos analisar outra base de dados e outra taxonomia que não fosse a do IEEE. As limitações encontradas podem ser consideradas também como propostas para trabalhos futuros.

As dificuldades encontradas foram em relação a subjetividade da criação de uma organização da informação, no nosso projeto, uma taxonomia. Mesmo com a amarração de uma metodologia a definição final de relacionamento entre áreas fica a cargo do autor ou do grupo. E com isso experiências, influências e outros fatores podem influenciar nas decisões.

Trabalhos Futuros

O trabalho mostra-se interessante e obteve resultados e avaliações positivos. Porém seus resultados foram para apenas uma base, o IEEE, e para a atualização de uma taxonomia desse instituto. Esse foi um dos primeiros passos para o auxílio da análise de redes sociais para uma melhoria efetiva numa organização do conhecimento e abre muitas portas para discussão de sua continuidade. Como trabalhos futuros podemos citar:

i. Utilizar as outras palavras-chave disponíveis pelo IEEE;
ii. Aplicação da metodologia em outras bases e organizações do conhecimento. (Ex. ACM, CNPq, DBLP etc)
iii. Como a metodologia é genérica pode ser aplicada em qualquer outra área. (Ex. Saúde, educação, ciências sociais etc.)
iv. Aprimorar a complexidade da organização da informação e a semântica, podendo evoluí-la para um tesauro e/ou uma ontologia. (Ex.: SKOS[6])
v. Trabalhar problemas de ambiguidade entre nomes de autores que podem aparecer;
vi. Utilizar outros tipos de sociogramas. (Ex.: Citação, Cocitação);
vii. Aumentar o número de avaliadores mais especialistas;

Referências Bibliográficas

Abilhoa, W. D. (2014). Um Método Para Extração De Palavras-Chave De Documentos Representados Em Grafos. Dissertação (Programa de Pós-Graduação (Stricto Sensu) em Engenharia Elétrica) - Universidade Presbiteriana Mackenzie - São Paulo.

Assis, J.; Moura, M. (2013). A norma ISO 25964 e a semântica latente das folksonomias: a interoperabilidade semântica em questão. In: II Congresso ISKO

[6] https://www.w3.org/2004/02/skos/

Brasil, 2013, Rio de Janeiro. Desafios e perspectivas científicas para a organização e representação do conhecimento na atualidade. Rio de Janeiro. v. 2. p. 212-217.

Azevedo, T. B. De, Vicente, M., e Rodriguez, R. Y. (2012). Análise do conhecimento com o uso das redes sociais.

Blondel, J.-L. Guillaume, R. Lambiotte, e E. Lefebvre, . (2008). "Fast unfolding of communities in large networks," J. Stat. Mech. Theory Exp., vol. 10008, no. 10, p. 6.

Castro, C. de M. (1985). Há produção científica no Brasil? Ciência e Cultura, São Paulo, v. 37, n. 7, p. 165-187, (supl.)

D. Blei, L. Carin e D. Dunson, (2010). "Probabilistic topic models", IEEE Signal Processing Magazine, vol. 27, no., pp. 55-65, 2010

D. Magatti, S. Calegari, D. Ciucci e F. Stella, (2009). "Automatic labeling of topics", ISDA 2009 – 9th International Conference on Intelligent Systems Design and Applications, pp. 1227-1232

Dias, T. M. R., e Moita, G. F, (2014). Caracterização e Análise De Redes De Palavras-Chave Em Repositórios De Publicações Científicas. II Encontro Mineiro de Modelagem Computacional.

Fernanda, P., e Leal, F. (2013). A interdisciplinaridade na pesquisa em Ontologias no Brasil : uma análise do evento ONTOBRAS a partir da coautoria e do acoplamento bibliográfico - edições 2010 , 2011 e 2012.

Freitas, J. L., Fátima, H. De, Silva, N., e Bufrem, L. S. (2012). Gestão do conhecimento e redes sociais: uma análise da literatura periódica científica da Ciência da Informação, 35–49.

J.H. Lau, D. Newman, S. Karimi e T. Baldwin, (2010). "Best Topic Word Selection for Topic Labelling",Methodology, no. 2010, pp. 605-613.

J.H. Lau, K. Grieser, D. Newman e T. Baldwin, (2011). "Automatic labeling of topic models",Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1536-1545

Kaur, J.; Gupta, V. (2010). Effective Approaches for Extraction of Keywords. International Journal of Computer Science Issues, p. 144-148.

Moura, M. (2009). Folksonomias, redes sociais e a formação para o tagging literacy: desafios para a organização da informação em ambientes colaborativos virtuais Inf. Inf., Londrina, v. 14, n. esp, p. 25 – 45.

Moura, M. (2011). Interoperabilidade semântica e ontologia semiótica: a construção e o compartilhamento de conceitos científicos em ambientes colaborativos online. Inf. Inf., Londrina, v. 16. n. 3. p. 165 – 179

Nolasco, Diogo, e Jonice Oliveira. (2016). "Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data." 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE.

Oliveira, D., Souza, F., Bottura, A., e Lima, M. (2013). Classificação das áreas de conhecimento do CNPq e o campo da Enfermagem : possibilidades e limites. Revista Brasileira de Enfermagem, 66, 60–65

Raghavan, U. N., Albert, R., e Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 76(3). http://doi.org/10.1103/PhysRevE.76.036106

Rose, S. et al. (2010). Automatic Keyword Extraction from Individual Documents. Text Mining: Applications and Theory.

Souza, R. F. de. (2006). Organização e representação de áreas do conhecimento em ciência e tecnologia: princípios de agregação em grandes áreas segundo diferentes contextos de produção e uso de informação. Enc. Bibli. Biblioteconomia e Ciência da Informação, Florianópolis, número especial, p.27-21.

Souza, R. F. de. (2006 b). Organização e representação do conhecimento no contexto da Ciência da Informação, da Comunicação Informação em Ciência e da Educação. Ed. UNIVALI, p. 111-125.

Wasserman, Stanley; Faust, Katherine. (1994). Social network analysis: methods and applications. Cambridge: Cambridge University Press, 825 p. (Structural analysis in the social sciences, v.8).