DOI: 10.5748/18CONTECSI/PSE/ESD/6797

XML DATA SIMILARITY ANALYSIS: A LITERATURE REVIEW

Reginalda Santos Silva - UNIVERSIDADE SALVADOR (UNIFACS) - Orcid: https://orcid.org/0000-0001-5409-1153

Adriana Maria Delgado - UNIVERSIDADE SALVADOR (UNIFACS) - Orcid: https://orcid.org/0000-0001-6291-9010

Paulo Caetano Da Silva - UNIVERSIDADE SALVADOR (UNIFACS) - Orcid: https://orcid.org/0000-0002-5038-2460

Rafael Neto Costa - UNIVERSIDADE SALVADOR (UNIFACS) - Orcid: https://orcid.org/0000-0002-7287-0692

This study aimed to identify similarity analysis techniques between data and XML documents through a literature review. This work is important because it allows us to identify the state of the art on techniques for analyzing the similarity of data and XML documents. A literature review was carried out using a methodology based on research questions, search strings, sources of scientific articles in the field of computing. Twelve works were identified that present techniques, methodologies and algorithms for XML similarity analysis. Identification of 12 proposals for similarity analysis of data and XML documents. The identified techniques allow the comparative analysis of data expressed in different XML document structures.

Keywords: XML similarity, Similarity analysis in XML data, Similarity analysis in XML documents, Similarity analysis, XML

ANÁLISE DE SIMILARIDADE EM DADOS XML: UMA REVISÃO DA LITERATURA

Este estudo teve como objetivo identificar técnicas de análise de similaridade entre dados e documentos XML por meio de uma revisão da literatura. Este trabalho é importante pois permite identificar o estado da arte sobre técnicas de análise de similaridade de dados e documentos XML. Foi realizada uma revisão da literatura usando uma metodologia baseada em questões de pesquisas, strings de buscas, fontes de artigos científicos na área de computação. Foram identificados 12 trabalhos que apresentam técnicas, metodologias e algoritmos para análise de similaridade XML. Identificação de 12 propostas para análise de similaridade de dados e documentos XML. As técnicas identificadas permitem a análise comparativa de dados expressos em diferentes estruturas de documentos XML.

Palavras-chave: Similaridade XML, Análise de similaridade em dados XML, Análise de similaridade em documentos XML, Análise de similaridade, XML

XML DATA SIMILARITY ANALYSIS: A LITERATURE REVIEW

Análise de Similaridade em Dados XML: Uma Revisão da Literatura

Reginalda Santos Silva - Universidade Salvador - reginalda_144@hotmail.com

Adriana Maria Pereira Delgado - Universidade Salvador - adrianampdelgado@gmail.com

Paulo Caetano Silva - Universidade Salvador - paulo.caetano@unifacs.br

Rafael Neto Costa - Universidade Salvador - neto.rnc20@gmail.com

Abstract: This article presents a literature review aiming to identify similarity analysis techniques for data represented in XML. Articles that addressed techniques to verify the similarity of XML were searched. During the research and registration process, several techniques were analyzed and compared with the results of other studies, but many of them were repeated in the selected articles, in others there was an improvement in the techniques used. The use of frameworks, the use of relational databases and other proposed techniques were observed. From the results the evaluations in search of similarity of the XML document with the use of algorithms were considered satisfactory, however, it was found that some proposals did not perform well in content analysis. This is due to a problem pointed out by some authors in the combination of contents, that is, when they referred to the use of different terms or names that denote the same entity or concept. It is expected that this work, based on the analysis of the proposed methodologies and techniques, will help organizations in the development of XML-based data models.

Keywords: XML similarity; Similarity analysis in XML data; Similarity analysis in XML documents; Similarity analysis; XML.

Resumo**:** Este artigo apresenta uma revisão da literatura com o objetivo de identificar técnicas de análise de similaridade de dados representados em XML. Foram pesquisados artigos que abordassem técnicas para verificar a similaridade de XML. Durante o processo de pesquisa e registro, diversas técnicas foram analisadas e comparadas com os resultados de outros estudos, porém muitas delas se repetiram nos artigos selecionados, em outros houve um aprimoramento nas técnicas utilizadas. Observou-se o uso de frameworks, o uso de bancos de dados relacionais e outras técnicas propostas. Pelos resultados pode-se perceber que as avaliações em busca de similaridade do documento XML com o uso de algoritmos, foram consideradas satisfatórias, porém, constatou-se que algumas propostas não obtiveram bom desempenho na análise de conteúdo. Isso se deve a um problema apontado por alguns autores na combinação dos conteúdos, ou seja, quando se referiram ao uso de diferentes termos ou nomes que denotam a mesma entidade ou conceito. Espera-se que este trabalho, com base na análise das metodologias e técnicas propostas, ajude as organizações no desenvolvimento de modelos de dados baseados em XML.

Palavras-Chaves: Similaridade XML; Análise de similaridade em dados XML; Análise de similaridade em documentos XML; Análise de similaridade; XML.

1. Introdução

Na década de 1990, o World Web Consortiun (W3C) iniciou um projeto de desenvolvimento de uma linguagem de marcação denominada Extensible Markup Language (XML), que tivesse semelhança à SGML (Standar Generalized Markup Language), e a facilidade de compreensão do HTML. Inicialmente a função do projeto era que a linguagem criada fosse capaz de ser lida por software, e se incorporasse às outras linguagens. Com a evolução e uso dessa tecnologia foi identificada a necessidade de desenvolver técnicas e metodologias, para estudos da análise de similaridade sobre os dados XML. Isto porque, em função da sua flexibilidade, a XML permite a representação de dados de diversas maneiras. Uma técnica bastante utilizada é a análise em SGBDs Relacionais, onde dados são armazenadas a partir da leitura de toda a estrutura do documento, e o mapeamento dos seus elementos para os campos da tabela, facilitando o armazenamento e o processamento dos dados.

A XML (Extensible Markup Language) define como os dados serão organizados em relação ao conteúdo de um documento, sua codificação utiliza marcadores. A vantagem da XML é a facilitação no compartilhamento de dados devido ao seu armazenamento ser em formato de texto, desta forma a leitura é feita por uma variedade de aplicativos, no qual podem ser feitas atualizações sem que informações importantes do documento sejam perdidas.

Por ser extensível, ter uma estrutura fácil de analisar e processar, a tecnologia XML tornou-se o formato padrão para a troca de dados entre aplicações Web, sendo assim, é necessário gerenciar os grandes volumes de dados atualmente existentes na Web de forma eficiente.

Estudos por metodologias de análise de similaridade XML identificam as diferenças de estrutura, conteúdo e semântica dos documentos XML. Uma das razões da heterogeneidade de representação de dados em XML, é, porque informações semelhantes podem ser representadas de diversas formas em XML, podem estar em um atributo de um elemento, podem estar em diferentes elementos que possuem nomes distintos, podem estar em estruturas XML diferentes. Segundo (Silva P C et al, p. 641-662, 2011) a heterogeneidade em dados XML pode ser classificada como sendo de, (I) sintaxe, na qual podem existir diversas formas de representações de conteúdo semântico iguais, e.g., representação da mesma informação em idiomas distintos, ou em diversas unidades de medidas (e.g. pés e metros); (II) semântica, na qual conceitos diferentes são representados por elementos de mesmo nome (e.g. vírus no domínio médico e no de informática); (III) estrutura, na qual as diferentes estruturas XML representam a mesma informação, (e.g. atributos ou elementos, elementos em distintos modelos de hierarquia). Näppilä et al descrevem sobre a importância na flexibilidade da representação em XML, no entato, transforma o uso de dados XML em um trabalho complexo. (Näppilä, Järvelin & Niemi, 2008). Devido a essas dificuldades encontradas foram propostas metodologias, técnicas e funções algorítmicas com a finalidade de avaliar as semelhanças de similaridade existente entre os documentos XML. Com o objetivo de catalogar e analisar as soluções propostas neste trabalho, foi realizada uma revisão da literatura.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a metodologia utilizada para a realização da revisão de pesquisa, Seção 3 descreve os resultados obtidos

nos artigos pesquisados, a Seção 4 descreve a análise da proposta dos artigos comparando as metodologias e técnicas utilizadas, e, Seção 5 a conclusão do trabalho.

2. Metodologia

Para o desenvolvimento deste trabalho utilizou-se a metodologia de pesquisa de acordo com a especificação (Lima; Mioto, 2007) e (Kitchenham; Charters, 2006). A pesquisa bibliográfica foi elaborada objetivando a revisão da literatura sobre similaridade de dados XML. A pesquisa bibliográfica teve como base material já organizado em livros, artigos científicos, teses e dissertações. Foi realizada uma revisão sistemática de literatura, cujas metas foram prover elementos para exposição dos dados da pesquisa, descobrir o mais adequado sistema para coleta e análise dos dados, conhecer os estudos disponíveis referente ao tema principal deste trabalho e colher informações fundamentais para a elaboração das possibilidades e das questões de pesquisas. Foi usado um protocolo de pesquisa (Quadro 2.1 e Quadro 2.2) baseado em palavras chaves (strings de busca) em português e inglês.

A formação dos termos (strings) para realizar buscas nos repositórios de artigos acadêmicos e científicos, foi estruturada na seguinte forma:

I. Os termos identificados, foram traduzidos para o inglês.

II. Execução da pesquisa de busca foi realizada a partir do termos inseridos nas *Strings* de buscas.

Quadro 2.1 *String* de buscas para o idioma português

Strings para buscas
XML, Similaridade XML, Semântica, Taxonomia. (“Taxonomia”, “Semântica”, “Metodologia” ou “Técnicas”)

Quadro 2.2 String de buscas para o idioma inglês

Strings para buscas
XML, XML Similarity, Semantics, Taxonomy. (“Taxonomy”, “Semantics”, “Methodology” or “Techniques”)

Para separação das fontes de pesquisa foram analisadas:

1. A possibilidade de examinar os artigos na web;
2. A existência de instrumento de busca aplicando palavras chaves;
3. A importância e destaque das fontes, considerando prioridade os publicados em congressos, revistas, artigos, dissertações e teses relacionadas aos tópicos detectados na pesquisa.

Foram usados os seguintes repositórios para pesquisa:

1. Google (http://www.google.com.br)
2. Google Scholar (http://scholar.google.com.br)
3. IEEEXplore (http://ieeexplore.ieee.org/Xplore/home.jsp)
4. Researchgate (https://www.researchgate.net/)
5. Scholar ((https://scholar.google.com.br/)

6. ACM (https://dl.acm.org/)
7. Smanticscholar (https://www.semanticscholar.org/)
8. Citeseerx (https://citeseerx.ist.psu.edu/index

A pesquisa bibliográfica teve como foco a realização de estudos por técnicas de similaridade XML quanto a estrutura, conteúdo, semântica, sintaxe e taxonomia em documentos XML.

Durante a pesquisa por busca de similaridade da XML, foram identificados 67 artigos, e 12 foram selecionados para análise e comparação em relação a outras técnicas e metodologias utilizadas, usamos como critério de seleção artigos mais atuais, entre os anos de 2015 e 2021, e de melhor desempenho de acordo com os estudos das técnicas e algoritmos usados citados pelos autores nos artigos científicos.

Como estudo de revisão adotamos parâmetros de inclusão e exclusão. Foi criado um mapa mental para auxiliar no processo de inclusão e exclusão dos artigos identificados nas buscas, conforme a (Figura 2.1 mapa mental). Os artigos foram classificados da seguinte forma: 0 = Difere do objetivo da pesquisa, 1 = O tema do artigo possui relação como tema secundário em relação ao objetivo da pesquisa, 2 = O tema principal do artigo é o mesmo do objetivo da pesquisa. Foram excluídos os artigos que não referenciam ou tratam o tema como tema secundário à proposta da pesquisa e artigos publicados em entre os anos de (2007 e 2014). Na inclusão foi verificado que vários artigos se repetiam, i.e. presentes, em quase todas as bases de buscas, comprovando a importância das análises dos estudos das técnicas de similaridade.

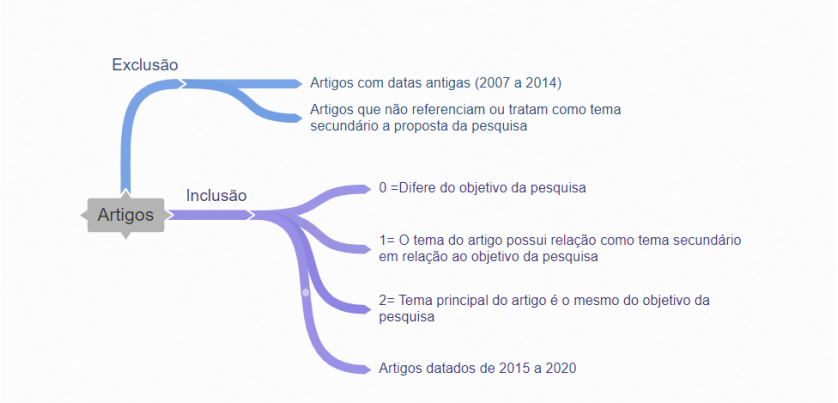

Figura 2.1 mapa mental

Foram catalogados os seguintes dados para registro dos trabalhos pré-selecionados:

I. Fonte
II. Nome do autor
III. Ano da Publicação
IV. Título do trabalho

3. Resultado

De acordo com os resultados das pesquisas, foram selecionados 12 artigos que serão explanados nesta seção. Serão descritas informações dos artigos selecionados conforme o uso de técnicas, algoritmos e metodologias de acordo a descrição dos autores sobre a utilização para o processo de análise de similaridade dos documentos XML.

3.1 Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents

Este trabalho é a continuidade de outro trabalho dos autores (Damny Magdaleno, Yadriel Miranda, Ivett E. Fuentes, María M. García, 2015). Nele é proposto uma metodologia para o OveralSIMsux, que é um cálculo usando algoritmos de agrupamento K-Start.

O algoritmo K-Start se propõe a capturar o grau de semelhança entre documentos, calculando uma matriz de similaridade (usando cosseno como medida) em uma função gerada através de uma equação conhecida como função de similaridade OVerallSimSUX. Nesse trabalho foi apresentada uma metodologia e comparado os algoritmos de agrupamento, Fuzzy-SKWIC, SKWIC, K-Start e o G-Start. E o resultado deste comparativo analisado concluiu que o algoritmo Fuzzy-SKWIC funciona melhor, embora não exista diferenças significativas em relação ao K-Start e G-Start. Portanto, fica a expectativa que no futuro surjam novos algoritmos de agrupamentos que venham ter efeitos mais significativos na medida de similaridade entre documentos XML.

3.2 An efficient similarity matching for clustering XML element

Saravadee Sae Tan e Gan Keng Hoon (2016) descrevem neste trabalho uma abordagem que calcula a similaridade em coleções de documentos XML com propósito de indicar a unidade de informação adequada para a tarefa de recuperação de dados. Deste modo, o processo utilizado é através da análise de clustering do XML que emprega medidas de similaridade em pares, desenvolvendo uma arquitetura para servir como unidade de recuperação no processo de correspondência por similaridade.

A arquitetura empregada no trabalho do processo de correspondência por similaridade divide-se em Análise XML e Aprendizagem semelhante. A análise XML deste trabalho define-se em um esquema elemento (ES) formando blocos de construção de elementos XML, e cada bloco possui seu nó-raiz e seus respectivos filhos. A aprendizagem semelhante define três medidas de similaridade chamadas de conteúdo, estrutural e híbrida. Nesse artigo, emprega-se a similaridade de conteúdo usando medida de distância Levenshtein, que se refere a duas strings pelo número mínimo de operações necessárias para transformar uma string em outra cadeia de caracteres. A distância N-gram é usada para prever probabilidades de correspondência entre strings, já a similaridade de cosseno utiliza-se de vetor N-gram.

Sendo assim, o trabalho apresenta uma correspondência de similaridade eficiente para agrupar elementos XML identificando a unidade de informação para a tarefa de recuperação de dados.

3.3 MXML: Implementation of a web-based application for merging XML documents using XML-SIM.

Este artigo propõe um aplicativo web, denominado MergerXML (MXML), que mescla os documentos XML **(**Viyanon, 2015). O MergerXML usa um método baseado na similaridade semântica do elemento XML, utilizando chaves denominadas XML-SIM (Viyanon and Madria, 2009). Estas chaves têm melhor desempenho na detecção da similaridade de falsos positivos e no tempo de execução quando comparado com o (XDoI (Viyanon, Madria e Bhowmick, 2008) que é a técnica para detecção da similaridade de subárvores agrupadas usando o nó-folha pai que se aproxima da correspondência da XML baseada no conteúdo e na estrutura, e o XDICSSK (Viyanon and Madria, 2009) são documentos XML agrupados usando os nós-folhas pais, o que produz um grande número

de subárvores, portanto é uma forma de aprimoramento da abordagem com o XDI-CSSK. O MXML integra recursos XML tanto em termos de estrutura, como de conteúdo semântico.

O autor descreve quatros etapas do funcionamento das buscas: (I) determinar subárvores para entidade de identidades de cada documento XML; (II) encontrar as chaves do nó-folha em que seu conteúdo é único; (III) encontrar chaves correspondentes de dois diferentes documentos XML e (IV) criar um documento a partir das correspondências utilizando o *Document Object Model* (DOM[1]) como uma interface de programação para documentos HTML, XML e SVG). Os autores descrevem o uso dos algoritmos (XDol, XDI-CSSK e XML-SIM), para detecção da similaridade de subárvores agrupadas usando o nó-folha pai para identificação da correspondência XML com base no conteúdo e na estrutura, o XDI-CSSK, que assim como o XDoI, agrupa os documentos XML usando nó-folha pais, o que vai produzir grande número de subárvores, por fim, o XML-SIM, segundo os autores melhor que XDoI e XDICSSK, pois descobre a semelhança do conteúdo. Esses algoritmos são usados para o desenvolvimento do MXML.

3.4 A hybrid method to evaluate XML Document similarity

Segundo (Yubiao Dai; Xueli Ren, 2016), o artigo tem como proposta descrever uma estrutura de documentos XML calculando a similaridade através do conjunto de caminho de correspondência difusos que compartilham semelhanças, e com isso, são gerados um método que classifica os documentos por categorias, com o objetivo da melhoria do desempenho do cálculo, desta forma examinam e definem o documento XML considerando a estrutura e semântica. Ademais, o artigo também referencia um algoritmo para calcular a correspondência de caminho ideal entre dois documentos, ou seja, calcula-se o valor médio entre dois documentos. Portanto, foram realizados experimentos utilizando os caminhos de correspondência de modo a mostrar o mais eficaz excluindo os caminhos duplicados que podem impactar na eficiência dos resultados.

3.5 Semantic Similarity of XML Documents Based on Structural and Content Analysis

De acordo com (Irvin Dongo; Regina Ticona-Herrera; Yudth Cadinale; Renato Guzmán, 2020), o artigo tem como proposta uma nova abordagem de Indexação Semântica Latente (LSI) que é uma técnica que recupera a informação do documento XML, e baseia-se em *Singular Value Dcomposition* (SVD), os autores descrevem que estendem o LSI que é composto na análise semântica da composição estrutural do documento XML, o qual para melhorar a medida de similaridade, calcula os seus termos como: os termos de estrutura, os termos do conteúdo, as ocorrências, sinonímia, polissemia. O algoritmo LSI adicionado ao contexto das palavras obtém melhor classificação de documentos comparado ao com a execução sem o contexto.

Para comprovação do desempenho da proposta foram realizados dois experimentos com bancos de dados, a realização dos experimentos obteve precisão quando a estrutura na análise está inclusa.

3.6 Mini-XML: An efficient mapping approach between XML and relational database

Segundo (Huchao Zhu; Huiqun Yu; Guisheng Fan; Huaiying Sun, 2017), o artigo tem como estudo a abordagem de mapeamento de banco de dados relacional, para

[1] https://www.w3.org/TR/dom41/.).

armazenamento de dados XML, com a abordagem do Mini-XML modelo de esquema de mapeamento, que usa a técnica baseada em caminho para mapear dados do documento XML para um banco de dados relacional, onde é feita uma comparação com o S-XML pois esta abordagem de armazenamento de dados semiestruturado, tem baixo armazenamento de dados em um banco relacional, e com isso, gerando aumento do tempo e do espaço de armazenamento.

Os autores concluíram que o Mini-XML comparado ao S-XML teve melhor desempenho em espaço e tempo, pois adota uma técnica baseada em caminhos como técnica básica entre os nós não-folha, rotulando os nós folha em sequência. E os resultados dos testes experimentais indicaram que, o Mini-XML consumiu menos espaço de armazenamento com a utilização do caminho do nó, sendo mais intuitiva e conveniente para a consulta em relação ao S-XML.

3.7 XS-Diff: XML schema change detection algorithm

Para (Abdullah Baqasah; Eric Pardede; Wenny Rahayu; Irena Holubovao, 2015) o XS-Diff é um algoritmo que utiliza a técnica de armazenamento de versões do esquema XML em banco de dados relacionais, no qual são identificados armazenamento de mudanças delta empregadas em tabelas relacionais. Faz uma comparação de desempenho em relação a outras ferramentas como: X-Diff proposto por Wang et al. (2003), XyDiff proposto por Cobena et al. (2002b), e DeltaXML é uma ferramenta comercial que fornece uma comparação para documentos XML e representa mudanças em XML. O XS-Diff que tem como propriedade estimar as mudanças ponderando a estrutura da árvore de esquema XML. Os autores descrevem que no estudo comprovaram a precisão do algoritmo avaliado no XSD sintético e reais, investigaram com o uso do X-Diff, XyDiff e DeltaXML e observaram que XS-Diff e XS-Diff, análogo ao X-Diff, é mais eficiente que o XyDiff. Analisaram também que tanto o X-Diff quanto o DeltaXML forneciam deltas que ficaram abaixo da marca ideal. O motivo é que as duas ferramentas são basicamente criadas para documentos XML, mas não para esquemas. No entanto o XS-Diff forneceu delta notável ou quase notável e os deltas derivados proporcionam a descrição da XML.

3.8 An efficient similarity-based approach for comparing XML documents

Segundo Alessandreia et al (2018), o artigo tem como proposta o Phoenix, abordagem utilizada como cálculo de similaridade recursiva para identificar partículas de correspondências entre os elementos do documento XML, sem necessitar o uso de chaves primárias. Para potencialização da comparação o Phoenix utiliza a programação dinâmica e algoritmos (XREL_CHANGE_SQL, XKeyDiff, XyDiff, XML-SIM-CHANGE) que comparam os mecanismos (por exemplo, nomes do elemento, atributos, conteúdo e subelemento) de documentos XML e calculando o grau de similaridade entre eles.

O X-Diff utiliza a assinatura de nó (o cálculo de um nó rash, juntamente com as assinaturas entre as versões dos filhos desse nó). Com a capacidade de trabalhar grandes quantidades de dados, pois funciona de forma ascendente na combinação e propagação dos nós de folha de correspondência para as subárvores acima. Como o uso do XREL_CHANGE_SQL, a implementação utiliza bancos de dados relacional para o armazenamento das versões XML, depois aplica a SQL para a revelação da similaridade. As técnicas como o XKeyDiff, XyDiff e XML-SIM-CHANGE, no entanto, no caso dos elementos que não tinham as chaves primárias ou valor dos elementos, verificou-se resultados ruins.

Phoenix é uma abordagem para o cálculo da similaridade recursiva, para verificar a revisão de similaridade da XML revelando as partículas que correspondem aos

elementos do documento XML. Os dois elementos raiz das revisões são comparados e dão prosseguimento de forma recursiva por mecanismos dos seus subelementos.

Para os autores a técnica de mineração de dados clustering tem como finalidade dividir um conjunto de dados em cluster. No decorrer dos experimentos, as correspondência com o uso do XyDiff nos elementos excluir e inserir fizeram as correspondências incorretas, entretanto com a atualização das chaves primárias pode chegar a uma alta precisão e recuperação, mas como o foco do contexto do artigo não cogitavam a presença de um esquema de chaves primária, então escolheram retirar o XyDiff das análise, enquanto o Phoenix teve melhor precisão mediana em todos os cenários, recuperando menos falsos positivos em suas correspondências. Os autores concluíram que o Phoenix tem a capacidade de achar as correspondências em um tempo de execução moderado de forma eficiente.

3.9 XChange A Semantic Diff Approach for XML Documents

Este trabalho apresenta a abordagem XChange que tem finalidade ajudar a identificar e compreender a semântica das mudanças ao analisar versões de um documento XML (Alessandreia et al, 2020). Sendo assim, o XChange utiliza-se de mecanismo de inferência baseado na linguagem de programação o Prolog (Programação Lógica) para analisar as modificações sintáticas nos atributos e elementos do documento, funciona também na identificação de elementos correspondente entre versões com o uso de correspondência por chave e correspondência por semelhança. Além disso, o XChange propõe a construção semiautomática de regras para o enriquecimento semântico com base na mineração de elementos que frequentemente são modificados juntos.

Por conseguinte, com o objetivo de provar a eficiência e eficácia do XChange foi realizado um estudo comparativo entre ele e o X-Diff (Wang et al. (2003), com o uso de usuários que comprovaram que o X-Diff tem um arquivo resultante menor, porém as tarefas concluídas do X-Diff foram demoradas, sujeito a erros e menos intuitivo.

3.10 Improved Centralized XML Query Processing Using Distributed Query Workload

Este artigo tem como proposta uma técnica de poda escolhida para o processamento de consulta distribuída, utilizando o DGReLab que é uma técnica de poda para processamento de consulta distribuída que utiliza ReLab+ (Samini et al, 2017) para criar rótulos de nó e caminhos diversos com o auxílio do DataGuide que realiza a indexação baseada em grafos (Samini et al, 2021). Segundo os autores, a técnica com o Data Guide usa uma versão simplificada da árvore XML para melhorar eficiência do processamento de consultas, onde a arquitetura é dividida em duas etapas: primeiro o pré-processamento, no qual se usa um analisador Simple APi for XML para verificar a boa formação do documento XML e, como segunda etapa, o processamento de consulta, onde os caminhos no documento XML forem únicos, sendo indexados usando rótulo do nó folha do caminho.

As consultas dos usuários são admitidas pelo processador de consulta global, onde são subdivididas em sub consultas, que detectam os processadores de consultas locais que abrangem as respostas parciais para as consultas. As sub consultas são processadas utilizando D-DGReLab+ DGReLab, e os resultados parciais são conduzidos a retornar ao servidor global para o processo de agrupamento. De acordo a analogia da apuração, o desempenho do D-DGReLab+, comparado com a técnica de processamento de consulta centralizada, DGReLab+, revelou que D-DGReLab+ obteve melhor execução para consultas intricadas e consultas com um grande número de nós de processamento.

3.11 Effective and Efficient XML Duplicate Detection Using Levenshtein Distance Algorithm

Gaikwad and Bogiri, (Gaikwad and Bogiri, 2015) fazem uma revisão da literatura, citando exemplos como: X-Diff expressado por (Yuan et al, 2003) como algoritmo de descoberta de transformação positiva para documento XML, ao qual contém a especialidade da XML para a metodologia de alteração de árvore para árvore. Esses autores descrevem o algoritmo para detecção entre as duas versões distintas de dados XML com ajuda de informações da estrutura XML. Discutem o DogmatiX, apresentado por (Herschel and Naumann, 2005), que é uma estrutura para detectar duplicata, que se apoia em três componentes que são, nomeadamente, definição candidata, definição duplicada e detecção de duplicados. A técnica DogmatiX verifica os elementos XML de acordo os valores diretos dos dados como também a estrutura dos pais filhos. Também discutem o um modelo sugerido por Zaiqing et al (Zaiqing et al, 2005) para identificar objetos de um domínio particular. No sistema proposto é usado o algoritmo de distância de Levenshtein, que utiliza uma rede bayesiana para estabelecer a possibilidade de dois objetos XML serem repetidos, e o modelo da Rede Bayesiana, no qual os algoritmos são misturados por estrutura dos objetos, de forma que, assim a possibilidade de todos os objetos é determinada levando em consideração não somente as informações que os objetos possuem, além das informações estruturadas. O algoritmo de distância de Levenshtein é bastante adaptável, o que implica na utilização de distintas similitudes medidas e no aspecto de estabelecer distintas possibilidades.

Usando o algoritmo de distância de Levenshtein e uma rede bayesiana para determinar a probabilidade de dois objetos XML serem duplicados, se possibilita maior flexibilidade e possibilita o uso de diferentes semelhanças e diferentes formas de combinar probabilidades, fornecendo bons resultados. Segundo os autores, tanto com os dados coletados artificialmente, como com dados reais, o algoritmo alcança alta precisão e com isso boas pontuações na detecção de documentos XML repetidos.

3.12 Structure Based XML Document Clustering a Review

Os autores têm como foco a técnica de agrupamento (*Clustering)* XML, tendo como base a estrutura XML. O método de *cluster* é uma atividade da mineração de dados sendo executado através de agrupamento de coleções de objetos que compartilham características semelhantes. Dong Huang et al (2020) define o clustering na mineração de dados como:

> "O clustering de dados é um problema fundamental na área de mineração de dados e aprendizado de máquina, cuja finalidade é particionar um conjunto de objetos em um determinado número de grupos homogêneos, cada um referido como um cluster".

Os autores descrevem algumas abordagens de *clustering* XML baseados em estrutura, para esclarecer que o *cluster* tem uma capacidade de desempenho superior ao processamento focado em conteúdo:

1. Abordagem com Definição de Tipo de Documento DTD (*Definition Type Document*) valida um modelo básico que contém a estrutura do documento como uma lista com determinadas propriedades e características. Os autores citam o Xcluster algoritmo de agrupamento de documentos XML, que utiliza a técnica de integração escalável para aumentar as fontes de dados e o XMine. Já o XMine modelo de esquema de mapeamento, usando técnica baseada em caminho para mapear dados do documento

XML em um banco de dados relacional. Sendo assim o DTD usa estruturas de árvores para simplificar as regras de transformação, ou seja, a medida de similaridade é comparada com os nós DTD examinados nos procedentes imediatos, desse modo os nós folhas geram informações de contesto para serem analisados e integrados como medidas globais;

2. Abordagem de similaridade de tag e caminho o documento é tratado como pacote de tags e o conteúdo do documento é desprezado. Essa abordagem é considerada como de baixa qualidade no agrupamento, já que desprezam as informações textuais e estruturais;
3. Editar abordagens de distância – fazem uso de representações de arvores convencionais de documentos XML, onde é medido a distância do número mínimo de procedimentos de operações fundamentais essências para modificar uma arvore em outra;
4. Abordagem de padrão – como abordagem de agrupamento utiliza o framework XProj. onde o documento na primeira etapa é divido subdivido aleatoriamente em divisões de tamanho idêntico para depois ser extraído e assim calcular a similaridade do documento. Sendo assim, o algoritmo repete o processo à medida que os documentos são verificados e estimados as interações e, portanto, o Xproj atinge um nível de precisão maior que outros métodos de distância de árvores.

4. Conclusão

Ao analisar os estudos dos 12 artigos descritos na Seção 3, foram verificadas diferentes tecnologias e metodologias juntamente com uso de algoritmos, para encontrar melhor desempenho de cálculo de similaridade nos documentos XML. Na grande maioria dos artigos as metodologias e técnicas envolviam estudos semânticos tanto de estrutura, como de conteúdo. Outros trabalhos usam técnicas de agrupamento associadas a diferentes algoritmos.

Muitas foram as técnicas de abordagens consideradas pelos autores como boas, e os autores indicaram novos estudos para aprimoramento daqueles em discussão. Na análise dos artigos estudados, percebeu-se que o MergerXML (MXML) revelou-se como sendo de grande desempenho, no que se refere a de tempo de execução, pois o documento XML é integrado em conteúdo semântico, como de estrutura. O MergerXML (MXML) mescla os documentos XML, esse método tem por base a similaridade semântica do elemento XML utilizando chaves denominadas XML-SIM.

A partir da Revisão da Literatura percebe-se que são poucos algoritmos e técnicas existentes para a análise de similaridade em XML, embora XML seja uma tecnologia bastante utilizada para a representação de dados, sua heterogeneidade de representação implica na necessidade de estudos sobre análise de similaridade estrutural e semântica na forma de representação de dados em XML;

Para continuidade deste trabalho mesclar as diferentes técnicas de forma que se possa conseguir uma solução mais eficiente que analise as diferentes estruturas XML, porém com a mesma semântica.

Referências Bibliográficas:

Baqasah, Abdullah *et al*. XSDiff: XML schema change detection algorithm, International Journal of Web and Grid Services, Computer Science, v.11, p. 160–1922, April 2015. Disponível em: https://www.researchgate.net/publication/275155963_XS-Diff_XML_schema_change_detection_algorithm. Acesso em: 17 de julho 2021. Doi: https://doi.org/10.1504/IJWGS.2015.068897.

Dong Huang et al (2020). Ultra-Scalable Spectral Clustering and Ensemble Clustering, Page(s): 1212 - 1226, Volume: 32, Issue: 6, 06 March 2019. Disponível em:

https://ieeexplore.ieee.org/document/8661522. Acesso em: 10 de agosto de 2021. Doi: 10.1109/TKDE.2019.2903410.

Gaikwad, Shital; Bogiri, Nagaraju. Effective and Efficient XML Duplicate Detection Using Levenshtein Distance Algorithm. Conferência Internacional sobre Computador, Comunicação e Controle (IC4), v. 4, Issue 6, jun. 2015. ISSN (Online): 2319-7064. Disponivel em: https://www.semanticscholar.org/paper/Effective-and-Efficient-XML-Duplicate-Detection-Gaikwad-Bogiri/2a448ca81162a166e8ebacf837e864639f87fb30. Acesso em: 17 de julho 2021.

Gaikwad, Shital; Bogiri, Nagaraju. Levenshtein distance algorithm for efficient and effective XML duplicate detection. International Conference on Computer, Communication and Control (IC4), set. 2015. Disponível em: https://ieeexplore.ieee.org/abstract/document/7375698. Acesso em: 17 de julho 2021. Doi: 10.1109/IC4.2015.7375698.

Herschel, M., & Naumann, F. (2005). DogmatiX tracks down duplicates in XML. SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, June 2005 Pages 431–442. Doi:10.1145/1066157.1066207.

Huchao, Zhu *et al*. Mini-XML: An efficient mapping approach between XML and relational database. International Conference on Computer and Information Science (ICIS), China, mai. 2017. Disponível em: https://ieeexplore.ieee.org/document/7960109. Acesso em: 17 de julho 2021. Doi: 10.1109/ICIS.2017.7960109.

Irvin Dongo; Regina Ticona-Herrera; Yudth Cadinale; Renato Guzmán. Semantic Similarity of XML Documents Based on Structural and Content Analy, International Symposium on Computer Science and Intelligent Control. November 17-19, 2020, Newcastle upon Tyne, UK. ACM, New York, NY, USA, 9 pages. Disponivel em: https://www.researchgate.net/publication/349331545_Semantic_Similarity_of_XML_Documents_Based_on_Structural_and_Content_Analysis. Acessado em: 02 de agosto de 2021. Doi:10.1145/3440084.3441185.

Kitchenham, Barbara; Charters, Stuart. Guidelines for performing Systematic Literature. Software Engineering (2007). Disponivel em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.471&rep=rep1&type=pdf. Acesso em: 17 de julho 2021.

Lima, Telma Cristiane Sasso de e Mioto, Regina Célia Tamaso. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. Revista Katálysis [online]. 2007, v. 10, n. spe, pp. 37-45. Disponível em: https://doi.org/10.1590/S1414-49802007000300004. Epub 25 Set 2007. ISSN 1982-0259.

Magdaleno, Damny et al. Clustering XML Documents Using Structure and Content based on a New Similarity Function OverallSimSUX. Computación y Sistemas, v. 19, n. 1 p. 151–161, mar. 2015. ISSN 2007-9737. Disponível em: https://www.researchgate.net/publication/282376658_Clustering_XML_Documents_Using_Structure_and_Content_based_on_a_New_Similarity_Function_OverallSimSUX. Acesso em: 17 de julho 2021. Doi: 10.13053/CyS-19-1-1922.

Magdaleno, Damny *et al*. Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. Revista Iberoamericana de Inteligência Artificial, Cuba, v.18, m. 55 p.1-11, jun. 2015. ISSN: 1137-3601. Disponível em: https://www.semanticscholar.org/paper/Comparative-Study-of-Clustering-Algorithms-using-Magdaleno-Miranda/2ce25e62b30b8022fe2f0be3ed4ebdbe35846622. Acesso em: 17 de julho 2021. Doi:10.4114/IA. V18I55.1097.

Oliveira, Alessandreia et al. An efficient similarity-based approach for comparing XML documents, 2018. Information Systems, v.7, p. 40-57, jul. 2018. Disponível em: https://www.researchgate.net/publication/326607437_An_Efficient_Similarity-based_Approach_for_Comparing_XML_Documents. Acesso em: 17 de julho 2021. Doi:10.1016/j.is.2018.07.001.

Oliveira, Alessandreia et al. XChange A Semantic Diff Approach for XML Documents, August 2020. Information Systems. Disponivel em: https://www.researchgate.net/publication/343380199_XChange_A_semantic_diff_approach_for_XML_documents. Acesso em: 09 de agosto de 2021. Doi: 10.1016 / j.is.2020.101610.

Piao, Yong; Tianyu, Wang. A tensor-based approach for calculating XML similarity. Conferência Internacional sobre Ciência da Computação e Tecnologia de Rede (ICCSNT), set. 2016. Disponível em: https://www.researchgate.net/publication/308222729_Tensor-based_approach_to_XML_similarity_calculation. Acesso em: 17 de julho 2021. Doi:10.13195/j.kzyjc.2015.0793.

Samini, Subramaniam; Su-Cheng, Haw; Lay-Ki, Soon. Improved Centralized XML Query Processing Using Distributed Query Workload, IEEE Access (IF 3.367), v.9, p.29127 - 29142, fev. 2012. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9351924. Acesso em: 17 de julho 2021. Doi:10.1109/ACCESS.2021.3058383.

Samini, Subramaniam; Su-Cheng, Haw; Lay-Ki, Soon, and K. L. Koong, ‘‘QTwig: A structural join algorithm for efficient query retrieval based on region-based labeling,’’ Int. J. Softw. Eng. Knowl. Eng., vol. 27, no. 2, pp. 321–342, 2017.

Silva, P. C., Cruz, M. S. H., & Times, V. C. (2011). XLDM: an Xlink-Based multidimensional metamodel. Journal of Information Systems and Technology Management, 8(3), 641-662. Doi: 10.4301/S1807-17752011000300007.

Tan, Saravadee Sae; Hoon, Gan Keng. An efficient similarity matching for clustering XML element. Conference on Information Retrieval and Knowledge Management (CAMP), Malásia, agu. 2016. Disponível em: https://ieeexplore.ieee.org/abstract/document/7806343. Acesso em: 17 de julho 2021. Doi:10.1109/INFRKM.2016.7806343.

Thulasi, A; Remya, KTV; Raju, G. Structure Based XML Document Clustering: A Review. International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Índia, dez. 2017. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8286068. Acesso em: 17 julho 2021. Doi: 10.1109/ICTUS.2017.8286068.

Viyanon, Waraporn. MXML: Implementation of a web-based application for merging XML documents using XML-SIM, 2015 13th International Conference on ICT and

Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015, pp. 5-10, doi: 10.1109/ICTKE.2015.7368462.

Viyanon, Waraporn and Madria, Sanjay Kumar. A System for Detecting Xml Similarity in Content and Structure Using Relational Database. Proceedings of the 18th ACM Conference on Information and Knowledge Management, pg. 1197–1206, Hong Kong, China, 2009. ISBN 9781605585123. Doi: 10.1145/1645953.1646105.

Viyanon W., Madria S.K. (2009) XML-SIM: Structure and Content Semantic Similarity Detection Using Keys. In: Meersman R., Dillon T., Herrero P. (eds) On the Move to Meaningful Internet Systems: OTM 2009. OTM 2009. Lecture Notes in Computer Science, vol 5871. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-05151-7_31.

Viyanon W., Madria S.K., Bhowmick S.S. (2008) XML Data Integration Based on Content and Structure Similarity Using Keys. In: Meersman R., Tari Z. (eds) On the Move to Meaningful Internet Systems: OTM 2008. OTM 2008. Lecture Notes in Computer Science, vol 5331. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88871-0_35.

Yubiao Dai; Xueli Ren. A hybrid method to evaluate XML Document similarity, International Conference on Education, Management, Computer and Society, Jan 2016. Disponível em: https://www.researchgate.net/publication/314683804_A_Hybrid_Method_to_Evaluate_Similarity_of_XML_Document. Acesso em: 21 de julho 2021. Doi:10.2991/emcs-16.2016.165.

Yuan Wang, David J. DeWitt, Jin yi Cai, X-Diff: An Effective Change Detection Algorithm for XML Documents, pp. 519-530, Umeshwar Dayal, Krithi Ramamritham, T. M. Vijayaraman (Ed.), 19th International Conference on Data Engineering, IEEE Computer Society Press, Bangalore, India, March 2003, 0-7803-7665-X. Disponivel em: https://ieeexplore.ieee.org/document/1260818. Acesso em: 09 de julho de 2021. Doi:10.1109/ICDE.2003.1260818.

Weis,Melanie; Felix Naumann. DogmatiX tracks down duplicates in XML, January 2005. Disponivel em: https://www.researchgate.net/publication/234830263_DogmatiX_tracks_down_duplicates_in_XML. Acesso em: 09 de julho de 2020. Doi:10.1145/1066157.1066207.

Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. 2005. Object-level ranking: bringing order to Web objects. In Proceedings of the 14th international conference on World Wide Web (WWW '05). Association for Computing Machinery, New York, NY, USA, 567–574. DOI: https://doi.org/10.1145/1060745.1060828.