ISSN - 2447-178X

EMPÍRICABR
Revista Brasileira de Gestão, Negócios e Tecnologia da Informação

MELHORIA DE PROCESSOS DE GESTÃO EM SAÚDE PÚBLICA: EXTRAÇÃO AUTOMÁTICA DE CONHECIMENTO E BUSCA SEMÂNTICA DE DOCUMENTOS NÃO ESTRUTURADOS

Cláudia Maria F. A. Ribeiro: claudia.ribeiro@ifrn.edu.br, Daniel Bastos: danielbastosan@gmail.com, Jamillo Santos: jamillo@gmail.com, Johann Guerra: johann.guerra200@gmail.com, Robinson Alves: robinson.alves@ifrn.edu.br

RESUMO

Este artigo trata da extração automática de conhecimento em documentos textuais, tema de relevância crescente para a melhoria de processos de gestão. Neste contexto, apresentamos a ferramenta XMeaning, que é um buscador semântico que visa facilitar a identificação de documentos relevantes, bem como sua aplicabilidade na área de saúde. A área de saúde é particularmente beneficiada por avanços em mineração de texto, dado o grande volume de documentos manipulados e as exigências legais para arquivamento por longos períodos. Assim, as discussões e resultados apresentados neste artigo refletem o exercício interdisciplinar da aplicação de modelos e técnicas computacionais sobre a área de gestão de processos, e busca enfatizar os benefícios mútuos advindos dessa aproximação.

PALAVRAS-CHAVE: Mineração de Texto, Gestão em Saúde, Busca Semântica

ABSTRACT

This paper deals with the automatic extraction of knowledge in textual documents, a subject of growing relevance for an improvement of management processes. In this context, we present the XMeaning tool, which is a semantic search engine that aims to facilitate the identification of relevant documents, as well as its application in the health area. In fact, the health care area is particularly beneficial from the advances in text mining, given the large volume of documents handled and the legal requirements for archiving for long periods. Thus, the discussions and results presented in the paper reflect the interdisciplinary exercise of the application of computational models and techniques on a process management area, and it seeks to emphasize the mutual benefits arising from the proposed approach.

KEYWORDS: Text Mining, Health Management, Semantic Search

1. INTRODUÇÃO

Nos últimos anos, organizações da área de saúde, tais como a OPAS - Organização Pan-Americana de Saúde e o SUS - Sistema Único de Saúde têm se esforçado para a melhoria contínua de seus processos internos, apoiando entidades que se ocupam em discutir assuntos de interesse para a saúde pública, por exemplo, as condições de trabalho de seus profissionais na América Latina. Tal posicionamento está em perfeito alinhamento com a crescente pressão por serviços de maior qualidade, o que vai além da qualidade intrínseca do serviço prestado, incluindo aspectos como atender em menor tempo e com menor custo.

Processos de gestão em saúde comumente usam grandes repositórios de documentos, em geral escritos em linguagem natural. A análise do conteúdo desses documentos é predominantemente manual, e por essa razão, extremamente sujeita a eventuais discordâncias e interpretações pessoais. Por exemplo, as Mesas de Negociação Permanente (MNPs), fórum criado para o acompanhamento de decisões relativas à melhoria das condições de trabalho no SUS, costumam registrar em atas as discussões sobre condições e reivindicações de trabalho dos profissionais de saúde.

Percebendo a importância de documentos para a melhoria dos processos de gestão em saúde, o Núcleo Avançado de Inovação Tecnológica (NAVI), do Campus Natal-Central (CNAT) do IFRN, em parceria com o Laboratório de Inovação em Saúde (LAIS) da UFRN, têm desenvolvido com o apoio do Ministério da Saúde, uma plataforma para o armazenamento de atas e demais documentos relevantes, originados nas Mesas de Negociação Permanente (MNPs). Apesar do avanço obtido pelo armazenamento nestas plataformas, foi observado que parte significativa dos documentos armazenados não possuem clareza e objetividade, nem estrutura padronizada, embora apresentem elementos textuais típicos de atas de reunião, tais como: participantes, data, local, pauta, discussão e encaminhamentos.

Documentos textuais, embora tenham valor como registro, a análise de seu conteúdo tem dependido da capacidade humana e seus limites em tratar grandes volumes. O processo automatizado de extração de conhecimento, portanto, representa um esforço para aumentar a eficiência e melhoria dos serviços de análise de documentos. É precisamente neste contexto que se insere este trabalho, que apresenta XMeaning, uma ferramenta de busca semântica, que se apoia em técnicas de mineração de texto. Esta ferramenta oferece uma forma inteligente de busca de documentos relevantes, a partir de algoritmos de similaridade semântica (MILLER; CHARLES, 1991).

Embora tenha sido aplicada à gestão em saúde, a ferramenta XMeaning também pode ser aplicável em outros domínios. Na verdade, a área de mineração de texto (Text Mining) tem produzido muitos algoritmos e soluções (TAN, 1999), comumente categorizados em algoritmos de agrupamento, categorização e recuperação (BERRY; CASTELLANOS, 2008), que vem sendo utilizados no auxílio à automatização de processos em diferentes áreas (MINER, 2012). Especificamente na área de saúde, muitas oportunidade e desafios têm fomentado a aplicação em mineração de texto (RAJA, 2008; KOH, 2011).

Aspectos técnicos específicos da ferramenta XMeaning bem como sua validação estão detalhados neste artigo, que está estruturado da seguinte forma. Esta introdução é seguida da seção

2, que é dedicada à discussão do estado da arte em mineração de texto.

As seções 3 e 4 tratam, respectivamente, da arquitetura e detalhamento das funcionalidades da ferramenta XMeaning. A seção 5 é dedicada ao detalhamento de experimentos e estudos de caso na área de saúde, estratégia de validação utilizada, bem como a análise dos resultados obtidos. Por fim, a seção 6 traz conclusões e perspectivas de trabalhos futuros.

2. MINERAÇÃO DE TEXTO

A Mineração de Texto é um campo interdisciplinar, que tem como objetivo principal descobrir e extrair conhecimento em documentos escritos em linguagem natural. A relevância desta área cresce conforme cresce a disponibilidade de documentos e a necessidade de processá-los automaticamente. A aplicabilidade das técnicas desenvolvidas nesta área vai além do tratamento de documentos, incluindo também o processamento de e-mails, blogs e redes sociais (AGGARWAL; ZHAI, 2012).

No segmento da saúde, aplicações que utilizam mineração de dados e texto podem beneficiar, de forma significativa, todas as partes envolvidas. Por exemplo, aplicações para detecção de fraude, reconhecimento prévio de pacientes de alto risco e aplicações para identificação de tratamentos mais efetivos, dentre outros, conforme reportado por TOMAR & AGARWAL (2013). As técnicas comumente utilizadas por esse tipo de aplicação, de forma geral, envolvem uma etapa de descoberta automatizada de conhecimento, também conhecida por KD (Knowledge Discovery). São duas as principais abordagens para KD: Descoberta de Conhecimento em Dados Estruturados (KDD) e a Descoberta de Conhecimento em Dados não Estruturados ou documentos textuais (KDT).

Diferentes denominações são usualmente associadas à descoberta automática de conhecimento, tais como Data Mining, Machine Learning e Statistical Learning (HOTHO; NÜRNBERGER; PAAß, 2005). Embora guardem similaridades, é importante distinguir estes termos. Data Mining é usualmente considerado sinônimo para KDD, sendo comumente aplicado para descoberta por meio de padrões, em grandes bancos de dados. Machine Learning (ML) é uma área da Inteligência Artificial (IA) relacionada ao desenvolvimento de técnicas que permitem aos computadores "aprenderem" a analisar conjuntos de dados. Por sua vez, Statistical Learning baseia-se na matemática, mais especificamente em métodos estatísticos para análise de conjunto de dados, sendo a probabilidade usada para modelar incerteza e aleatoriedade. Assim, embora sejam comunidades que priorizam o desenvolvimento técnicas específicas, tais técnicas podem ser utilizadas conjuntamente, uma vez que o foco é a extração automática de conhecimento.

O volume a ser tratado não se constitui o único desafio da descoberta automática de conhecimento. A heterogeneidade dos dados, e em especial a subjetividade das linguagens naturais, têm impulsionado a busca por algoritmos mais eficientes. Por exemplo, Deep Learning surgiu como uma nova área de pesquisa em Machine Learning, que investiga como o funcionamento do cérebro humano pode ser modelado computacionalmente, e assim criar mecanismos mais eficientes

para reconhecimento de imagens e processamento de linguagem natural (PLN). De fato, estudos apontam que aproximadamente 80% da informação contida na web se apresenta em forma de texto, ou como costuma ser denominado, documentos não-estruturados (GANTZ; REINSEL, 2012).

A ferramenta XMeaning descrita neste documento, portanto, pretende ser instrumento que auxilia a extração automática de conhecimento em documentos textuais não-estruturados, a fim de auxiliar os processos de gestão. Inicialmente aplicado no contexto dos processos de gestão em saúde, a ferramenta XMeaning pode ser aplicado a qualquer processo que envolva a manipulação de documentos em linguagem natural.

Estas normas têm como objetivo dar uma orientação geral aos autores dos artigos no momento em que forem redigir e, principalmente, quando forem organizar e digitar seus artigos científicos.

3. FERRAMENTA XMEANING

3.1. Visão geral

A concepção da ferramenta XMeaning está em conformidade com a arquitetura orientada a serviço, também conhecida por SOA (do inglês Service Oriented Architecture). Os princípios fundamentais da orientação a serviço incluem, dentre outros: (1) Baixo acoplamento, relacionado à capacidade de ser independente de outros serviços para realizar a sua tarefa; (2) Interoperabilidade, que permite a comunicação entre serviços desenvolvidos em diferentes tecnologias; (3) Reusabilidade e composição, que permite o desenvolvimento de serviços complexos, a partir das funcionalidades de outros serviços (VALIPOUR, 2009).

No contexto dos sistemas de saúde, os serviços web (Web Services) são particularmente interessantes, uma vez que podem facilmente ser integrados aos sistemas já existentes. Assim, a ferramenta XMeaning pode ser integrada tanto ao sistema de Mesas de Negociação, quanto outros sistemas que manipulem documentos textuais não estruturados. A Figura 1 detalha a arquitetura da ferramenta XMeaning, onde é possível identificar dois módulos principais, um módulo de pré-processamento e um módulo de extração de conhecimento.

O módulo de pré-processamento prevê a manipulação de documentos textuais estruturados e não estruturados. É importante distinguir a natureza da estrutura considerada, que difere do conceito tradicionalmente associado a bases de dados, que são objetos de ferramentas especificas KDD, conforme discutido anteriormente. Neste contexto, considera-se estrutura um formato de texto já conhecido, por exemplo, atas, memorandos, artigos, etc. A identificação prévia de uma estrutura permite tornar o algoritmo de extração mais eficiente e rápido.

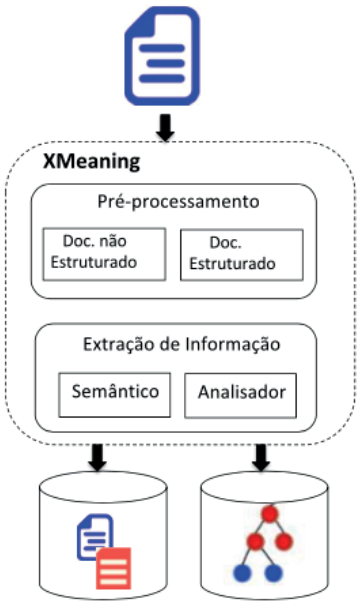

Figura 1. Visão geral da ferramenta XMeaning

Na base das Mesas de Negociação, as atas são documentos preponderantes, mas foi observada a ausência de padronização, com diferentes formatos adotados pelas mesas de negociação. Tal divergência na estrutura dos documentos, contudo, não impede a utilização da ferramenta XMeaning. Neste caso, o analisador é um módulo específico que extrai a estrutura do documento, qual seja. De fato, neste caso o que se pretende é obter a pauta e o texto que descreve a discussão e os encaminhamentos. Além desses, outros elementos mais facilmente identificáveis também são tratados, tais como a data de realização da reunião e os participantes.

O segundo módulo é responsável pela extração de conhecimento, propriamente dita. Este módulo é composto de dois submódulos. O primeiro é um analisador, componente que recebe o documento pré-processado e analisa os elementos textuais individualmente. Por exemplo, extraindo data e local de realização, participantes e itens da pauta, dentre outros. A parte referente a discussão e encaminhamento é que apresenta o maior desafio, mas é exatamente o elemento textual mais rico e de interesse para extração de conhecimento relevante.

O tratamento feito pelo analisador permite desconsiderar elementos textuais menos relevantes, para que os termos de maior relevância, como substantivos e adjetivos sejam extraídos. Foi considerado um limite de 10 termos relevantes, que são armazenados em formato de metadado para descrever o documento em questão, que também é armazenado na íntegra. O componente semântico permite ampliar a busca por termos similares ou formas mais relevantes, a partir de termos encontrados no documento.

Para fins de ilustração e melhor compreensão sobre o funcionamento da ferramenta XMeaning, consideremos que as condições de trabalho dos profissionais envolvidos em epidemias e surtos, como dengue, seja objeto de discussão de diversas Mesas de Negociação. Neste caso, é interessante

para uma Mesa, por exemplo, obter atas e demais documentos de outras mesas, a fim de avaliar os avanços nesta discussão. Esta busca é precisamente uma outra funcionalidade da ferramenta XMeaning, baseada na similaridade semântica, tomando como base os metadados.

3.2. Aspectos tecnológicos

A ferramenta XMeaning foi desenvolvida para ser facilmente integrável a outros componentes de software e segue orientações e práticas comuns ao desenvolvimento de aplicações distribuídas. Assim, por meio da Figura 2 é possível distinguir componentes de front-end, como é comumente chamado o lado cliente da aplicação, a partir do qual as funcionalidades são solicitadas.

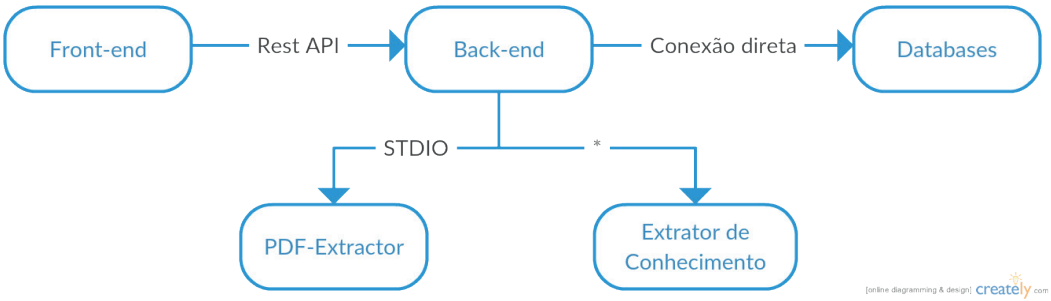

Figura 2 - Visão tecnológica da ferramenta XMeaning

As funcionalidades são disponibilizadas a partir do back-end, que é a parte servidora da aplicação, que tem dentre outras atribuições, disciplinar e proteger o acesso a bases de dados. No caso da ferramenta XMeaning, fazem parte do back-end os módulos responsáveis pelo pré-processamento (PDF-Extractor) e pela extração de conhecimento.

A comunicação entre o front-end e o back-end é feito por meio de uma API (Application Programming Interface). Este tipo de componente arquitetural é fundamental para obter o engajamento, visando a integração plena entre aplicações. APIs materializam o conceito de desenvolvimento baseado em contrato (MILANOVIC, 2005), que permite um desenvolvimento agnóstico, independente de tecnologias específicas. Assim, uma aplicação que queira utilizar ou consumir as funcionalidades da ferramenta XMeaning, basta acessar a API sem precisar conhecer nem adotar as tecnologias que foram utilizadas em seu desenvolvimento.

A comunicação entre os componentes de uma aplicação distribuída é um dos aspectos técnicos de maior relevância para a arquitetura de software. Dentre os diferentes mecanismos de comunicação distribuída merecem destaque os Web Services, em especial a implementação baseada em REST (CHRISTENSEN, 2009). Este mecanismo é de fácil implementação e permite aproveitar todos os recursos já utilizados na navegação na Web, ou seja, na comunicação entre os navegadores (chamados clientes HTTP) e os servidores Web (chamados servidores HTTP).

A comunicação com o banco de dados é feita pelo back-end, o que confere a aplicação um maior isolamento e segurança a acessos não autorizados. Embora, atas sejam em geral, documentos de acesso público, essa característica permite utilizar a ferramenta XMeaning também para documentos

privados. A biblioteca STDIO foi utilizada para manipulação de documentos não estruturados em PDF, na fase de pré-processamento, quanto na fase de Extração de Conhecimento, conforme descrito na seção 3.1 e visualizado por meio da Figura 1.

4. MINERAÇÃO DE TEXTO COM XMEANING

4.1. Análise preliminar

O processamento de linguagem natural apresenta muitos desafios, principalmente em documentos não-estruturados. No caso de atas de reunião, documento-alvo examinado neste projeto, existem dificuldades adicionais, tais como ambiguidade na escrita e a multiplicidade de assuntos que podem ser tratados durante cada reunião, a utilização de diferentes siglas, dentre outros.

Algumas estratégias foram adotadas para tratar a complexidade inerente a este tipo de documento. Primeiro foi feita uma análise baseada na extração manual de um conjunto mínimo de termos, que pudessem descrever um determinado documento. Nesta etapa, foram analisadas 3 atas de diferentes regiões do país. Cada ata foi então submetida a extração por duas ferramentas de extração semântica existentes: Dandelion e Alchemy.

O objetivo desta fase foi efetuar uma análise comparativa da eficiência das duas ferramentas na extração de termos relevantes. Os resultados desta análise são ilustrados pela Figura 3, a seguir. Em cada gráfico podem ser visualizadas a quantidade de termos extraídos pelo Dandelion, pelo Alchemy e os termos que foram identificados por ambas as ferramentas.

Os resultados obtidos demonstraram a irregularidade na extração de termos entre atas, tanto em termos quantitativos, quanto qualitativos. Na análise qualitativa observou-se o conjunto de termos que foram identificados por ambas as ferramentas. Além disso, foi realizada uma análise qualitativa em relação à extração manual realizada inicialmente.

Tal análise deixou explícita a ineficiência da utilização direta destas ferramentas para extração semântica de termos, a partir do tipo de documento-alvo. Dessa forma, uma arquitetura híbrida foi concebida, o que deu origem a ferramenta XMeaning.

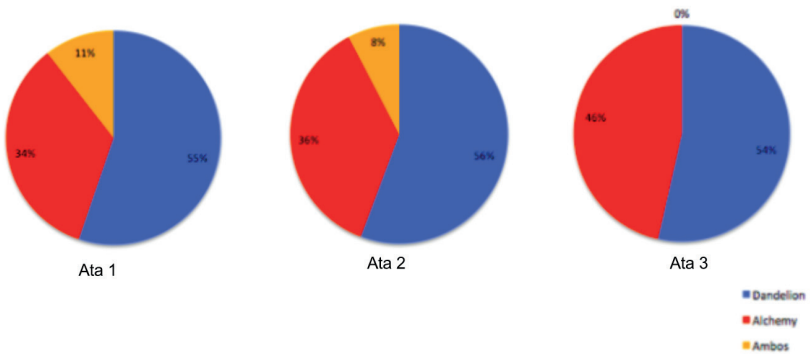

Figura 3 - Análise comparativa das ferramentas Dandelion e Alchemy

4.2. Aspectos operacionais

Considerando os repositórios de documentos das mesas de negociação, a ferramenta XMeaning pode ser ativada durante o processo de inclusão de uma ata no repositório. Os passos realizados podem ser visualizados na Figura 4.

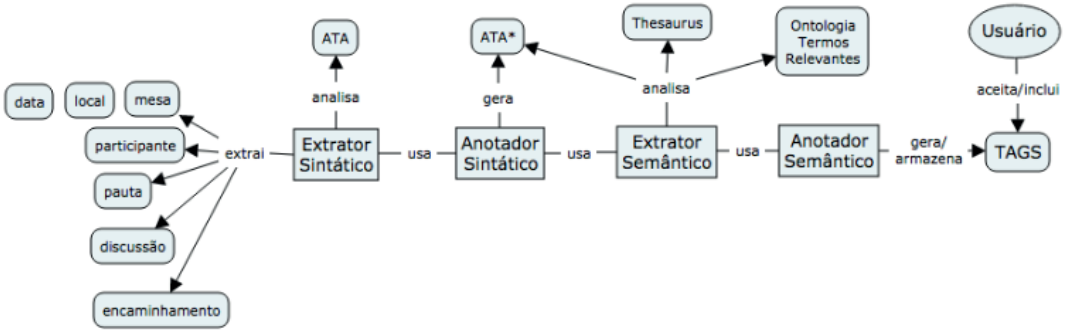

Figura 5 – Visão geral XMeaning

Figura 4 - Fluxo operacional da inclusão de uma ata com a ferramenta XMeaning

O Extrator sintático, como o próprio nome indica, extrai elementos textuais como parágrafos e frases, descartando elementos que não carregam significado, classificando as palavras em categorias, tais como substantivos, adjetivos e verbos.

Além disso, durante a inclusão (upload) do documento, o extrator sintático efetua uma análise com o objetivo de identificar elementos textuais relevantes. Considerando que esse tipo de documento, em geral, é o resultado concreto de uma reunião entre atores interessados na discussão de um tema específico, e que costuma resultar em encaminhamentos, o extrator sintático é treinado para descobrir tais elementos, que formam um modelo conceitual conforme Figura 5, a seguir.

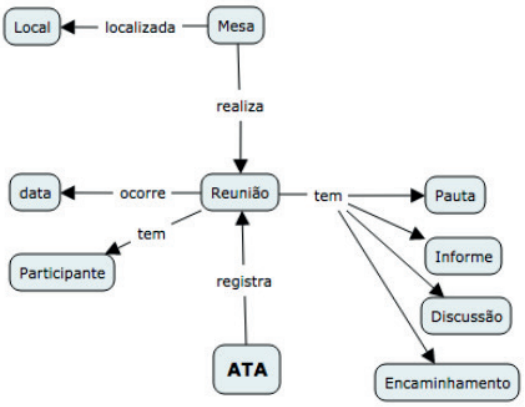

Figura 5 - Modelo Conceitual de uma ata

Seguindo o fluxo, o Anotador sintático gera a seguir, um documento semiestruturado (A T A*), conforme os elementos de interesse, conforme Figura 5: data, local, participantes, pauta, discussão e encaminhamentos.

No terceiro passo do fluxo, observa-se o Extrator semântico, componente que estende a análise da ata, visando identificar um conjunto mínimo de conceitos (TAGs), a partir dos quais uma ata pode ser resumida. Por exemplo, em uma reunião para tratar de condições especiais de trabalho para a equipe de saúde, durante uma epidemia de dengue, o conjunto mínimo de Tags poderia incluir: dengue, epidemia, periculosidade, adicional, equipe de saúde, além de outros termos de relevância, que possam ser utilizados para descrever brevemente o documento e posteriormente localizá-lo.

O extrator semântico, além da ata semiestruturada (A T A*), também utiliza dois componentes externos, um thesaurus e uma ontologia de termos relevantes às mesas de negociação. Importante considerar que a ontologia, com seus conceitos e relacionamento entre os mesmos, foi fornecida por um especialista. Estes componentes possibilitam maior qualidade a descrição do documento, podendo inclusive, suprir eventuais erros e problemas de baixa qualidade textual do documento.

Uma vez que um conjunto de conceitos mínimos é identificado, o componente Anotador semântico é o responsável por gerar Tags, que após validados pelo usuário serão associados a ata correspondente. Esta etapa permite que o usuário interaja com a ferramenta XMeaning, conferindo-lhe maior precisão. Embora ainda não esteja disponível, pretende-se que esta etapa seja estendida para alimentar a ontologia, ampliando o vocabulário com os termos mais relevantes, gerando uma base de conhecimento alimentada automaticamente. Tal funcionalidade deve ser melhor investigada, de forma a evitar a poluição da base pela inclusão desordenada de termos.

4.3. Extração de termos-chave

A Extração de Palavras Chave (KE, do inglês Keyword Extraction) é a tarefa de identificação automática de um conjunto de termos que descreva satisfatoriamente o assunto de um documento. Tal função é de suma importância para o estudo de Processamento de Linguagem Natural, uma vez que permite a execução de diversas tarefas, como Extração de Informação, Mineração de Texto e Categorização de Texto (BELIGA; MEŠTROVIĆ; MARTINČIĆ-IPŠIĆ, 2015).

Neste projeto, a Extração de Palavras Chave é utilizada extensivamente na indexação dos documentos, tendo em vista a busca baseada nos termos extraídos. No contexto aqui proposto, estes termos serão denominados TAGs, e cada documento será indexado de acordo com um conjunto de no máximo dez tags, denominado tagset.

No presente trabalho, descreveremos a experiência de aplicação de dois algoritmos para a realização do processo de KE. Estes algoritmos pertencem a duas classes diferentes de métodos de extração, a saber, a) métodos estatísticos, aqueles que extraem as tags através da contagem do número de ocorrência de um termo em determinado contexto; e b) métodos estáticos, que filtram os termos presentes em um documento de acordo com uma lista pré-definida de palavras que podem ser utilizadas como tags. Além disso, investigamos o uso de duas APIs externas dedicadas ao processamento de textos.

4.3.1 Método estatístico

Os métodos estatísticos de KE são algoritmos que executam a extração de tags a partir de técnicas de contagem do número de ocorrências de um termo em um conjunto de documentos. Neste projeto, foi utilizado um dos métodos mais antigos presentes na literatura atual, o Tf-Idf. Apesar de antigo, este algoritmo apresenta alto nível de precisão, quando comparado a outros mais recentes, como investigado em (BELIGA; MEŠTROVIĆ; MARTINČIĆ-IPŠIĆ, 2009).

O Tf-Idf (do inglês Term Frequency - Inverse Document Frequency, ou Frequência do Termo - Frequência Inversa do Documento) funciona a partir da execução de duas funções para cada termo do documento: a primeira função (Tf) retorna um número que representa a frequência do termo naquele documento; a segunda (Idf), retorna um número que representa a frequência do termo em todos os documentos na base.

O score de um termo δ é obtido dividindo-se Tf(δ) por Idf(δ). Para a nossa aplicação, foi utilizada a implementação do Tf-Idf presente na biblioteca Natural. Nessa implementação, a função Tf de δ é uma contagem do número de vezes em que o termo δ aparece no documento atual. Já a função Idf de δ conta o número de documentos no qual o termo δ está presente. A extração das palavras-chaves se dá com a ordenação dos termos em ordem decrescente de score, seguida da escolha dos 10 primeiros termos.

4.3.2. Método estático

Um método de extração é considerado estático quando realiza a extração através do uso de uma lista pré-definida de palavras (um Gazetteer) que podem ser escolhidas como tags de um documento adicionado. Para desenvolvê-lo, um conjunto de palavras essenciais ao escopo da ferramenta XMeaning foi obtido através da consultoria de um especialista no assunto. Utilizando uma versão em português da WordNet, chamada OntoPT foi possível obter uma lista de sinônimos, hipônimos e palavras similares às presentes no conjunto original. Essa lista tornou-se o nosso Gazetteer (RODRIGUEZ-MURO; KONTCHAKOV; ZAKHARYASCHEV, 2013). Assim sendo, a extração segundo esse método baseia-se na busca dos termos presente no Gazetteer, em cada um dos documentos inseridos.

4.3.3. **APIs externas**

Além de implementações dos algoritmos citados nas subseções anteriores, também foram utilizados serviços externos disponíveis via APIs REST, para a execução da extração de termos-chave: Alchemy API e Dandelion.

A Alchemy é uma companhia subsidiária da IBM que presta serviços de Inteligência Artificial através de uma API REST extensivamente documentada. Os serviços oferecidos na área de Processamento de Linguagem Natural para textos em português variam desde o reconhecimento do autor de um documento até a extração de conceitos citados no texto e as relações entre eles. É importante citar que, apesar do alto nível de documentação quanto ao uso do serviço de Keyword Extraction utilizando a Alchemy API, há uma quantidade extremamente limitada de informações

públicas quanto ao funcionamento interno desta API, de forma que é impossível se fazer quaisquer conjecturas sobre os algoritmos usados para a prestação deste serviço.

O Dandelion é um produto da empresa italiana SpazioData, cujo objetivo é a extração de informações a partir de textos em diversos idiomas, entre ele o Português. Apesar de possuir uma gama menor de serviços do que o Alchemy, o Dandelion oferece suporte às tarefas necessárias. Assim com a IBM, a SpazioData não oferece nenhum detalhamento quanto ao funcionamento interno de suas APIs.

4.3.4. Considerações sobre escolhas tecnológicas

Após experimentos de validações, foram efetuadas algumas escolhas tecnológicas para o aperfeiçoamento do desenvolvimento da ferramenta XMeaning. O algoritmo Tf-Idf foi escolhido, por ter apresentado resultados muito superiores aos concorrentes.

Com relação as APIs externas, foram consideradas as métricas Precision e Recall, padrões de facto da comunidade científica para avaliação deste tipo de ferramenta. O Dandelion, candidato mais bem avaliado, apresentou Recall de apenas 35,9% no conjunto de atas utilizado, e Precision de 19.44%. Isso significa que, no caso de um usuário adicionar uma ata no sistema e 10 tags serem extraídas automaticamente pelo Dandelion, uma média de 8 dessas tags teriam de ser corrigidas manualmente.

O Alchemy API, apresentou resultados muito semelhantes ao do Dandelion. O seu Recall foi de 18,06% e a Precision foi de 19,44%. Devido à falta de informações públicas disponíveis sobre o funcionamento interno dessas ferramentas, é impossível tecer quaisquer conjecturas a respeito do motivo de um desempenho tão abaixo do esperado nos testes realizados.

Durante a análise, observou-se que, para um mesmo texto, o Dandelion retornava múltiplas instâncias do mesmo conceito (por exemplo, em uma mesma ata, quatro sinônimos da palavra “Bahia” foram retornados), o que tornaria o processo de utilização do sistema extremamente lento, devido à necessidade de remoção manual das tags duplicadas por parte do usuário. Em contrapartida, o Dandelion se mostrou extremamente eficaz na detecção de entidades nomeadas (Named Entities), como localidades, pessoas, e organizações. Aproximadamente 90% dessas entidades foram retornadas como tags.

Após análise dos resultados, foi averiguado que tanto o Dandelion quanto o Alchemy API apresentam desempenhos muito abaixo do esperado, e que o uso de qualquer um dos dois como solução final para o nosso projeto acarretaria em uma experiência de uso extremamente frustrante para os usuários do sistema.

O Algoritmo de Extração Estática (ver seção 4.3.2) também apresentou um resultado muito abaixo do esperado. Para esse método, a Precision apurada foi de 15,34%. Acreditamos que baixo desempenho apresentado seja decorrente do fato de que a lista de palavras utilizadas no Gazetteer é fixa.

Consequentemente, qualquer termo significativo que esteja presente em um texto, mas não conste no Gazetteer é automaticamente ignorada pelo algoritmo. O Tf-Idf apresentou Precision de 64,32%. Além do desempenho muito superior aos outros métodos de extração testados, o Tf-Idf possui as vantagens de ser um algoritmo altamente reutilizável, já que não exige a preparação de um Gazetteer,

e de ter sido amplamente testado pela comunidade científica, como demonstrado por HASAN & NG (2010). A Tabela 1 demonstra os resultados da análise.

Tabela 1 - Análise comparativa entre métodos de extração de termo-chave

Algoritmo	Precision	Recall
Tf-Idf	64.32%	N/A
Dandelion	19.44%	35.90%
Alchemy	19.44%	18.06%
Método Estático	15.34%	N/A

Importante pontuar que durante a execução desta análise, foi observado que todas as tags retornadas pelo Tf-Idf eram compostas de apenas uma palavra. Assim sendo, tags pertinentes a muitas das atas utilizadas, como “plano de carreira”, foram divididas pelo algoritmo em várias tags — no caso, “plano” e “carreira”. Tal comportamento é extremamente indesejável no contexto em questão, assim foi adicionada uma fase de pré-processamento que extraia sintagmas nominais dos documentos. Os sintagmas nominais são conjuntos de palavras que exercem a função de substantivo em uma frase. Uma vez extraídos, esses elementos podem ser utilizados como parâmetros de entrada do Tf-Idf, dando prosseguimento ao processo segundo o modelo apresentado anteriormente para esse algoritmo.

5. ARMAZENAMENTO

Os componentes arquiteturais da ferramenta XMeaning, discutidos anteriormente, têm como real função preparar para uma busca mais inteligente, também chamada busca semântica. Nesse contexto, foi fundamental incluir um banco de dados apropriado a este desafio. A escolha recaiu sobre o Neo4j, um banco de dados cujo modelo de armazenamento e busca de informações é totalmente baseado em grafos.

Para a ferramenta XMeaning cada conceito pertinente ao campo de conhecimento abordado nos documentos adicionados será armazenado como um vértice no grafo do Neo4j, e suas relações serão armazenados como arestas do mesmo. Além disso, as atas adicionadas ao repositório, juntamente com os metadados de cada uma delas, são representadas por vértices do mesmo grafo, e são ligadas às suas tags através de arestas. A manipulação de todas essas informações dá-se através da linguagem de consulta de dados Cypher e de uma API fornecida pelo próprio Neo4J.

6. BUSCA SEMÂNTICA

O principal objetivo da ferramenta XMeaning é, de fato, permitir a realização de buscas em um repositório de documentos, levando em consideração relações semânticas. Tais relações, embora

óbvias para o observador humano, não são compreensíveis para sistemas computacionais que realizam buscas através dos métodos convencionais. Assim, é que a interface é voltada para essa busca, conforme visualizada na Figura 6.

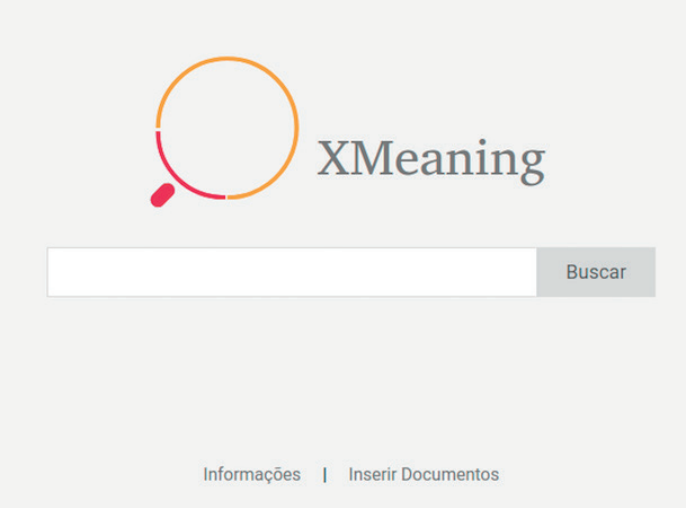

Figura 6 - Tela inicial da ferramenta XMeaning

Por esta interface, o usuário pode inserir documentos no repositório, procedimento que ativa todas as fases detalhadas no fluxo detalhado na seção 4. Após o usuário selecionar a opção “Inserir Documentos” e informar o arquivo, o extrator sintático da ferramenta XMeaning inicia sua atividade, assim como os demais componentes. O resultado desse fluxo pode ser visualizado na Figura 7.

Dentre as informações extraídas, estão incluídos o título, a data de publicação, a pauta, o texto da discussão e o conjunto de tags. Caso a ferramenta falhe em obter estas informações, o usuário pode incluir e salvar. Com relação as tags, o usuário pode aceitar, excluir e incluir tags que melhor descrevam o documento.

Importante observar que quanto melhor for redigida a ata, melhor será a extração automática. Nesse sentido, acredita-se que a ferramenta XMeaning também traz como benefício, a melhoria gradativa da elaboração das atas e outros documentos, conferindo maior qualidade aos repositórios de documentos.

ISSN - 2447-178X

Figura 7 - Resultado da extração automática

A busca semântica da ferramenta XMeaning, conforme mencionado anteriormente, baseia-se no conjunto de tags, a partir do qual, é possível relacionar diferentes documentos no repositório. A Figura 8, a seguir, exemplifica o resultado de busca pela tag “processo”, a partir do qual, dois documentos são recuperados.

Figura 8 - Busca semântica da ferramenta XMeaning

A interface da ferramenta XMeaning é propositalmente simples, de forma a privilegiar a função de busca semântica, embora possa ser facilmente customizada para ser melhor adaptada às plataformas ou aplicações as quais, XMeaning é inserida como funcionalidade adicional.

7. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho dedicou-se a explorar, em um caso prático, a interdisciplinaridade entre as áreas de computação e gestão, por meio do desenvolvimento da ferramenta XMeaning e sua validação na área de gestão em saúde pública. A capacidade de manipulação de documentos em grandes repositórios visa colaborar com os processos de gestão em saúde, que se caracteriza pelo grande volume de documentos escritos em linguagem natural, o que torna a busca e análise manual do conteúdo uma tarefa maçante e custosa.

A falta de estruturação e a complexidade da linguagem natural consistem em grandes desafios à extração automática de conhecimento em documentos textuais, porém os benefícios são claros. Além de permitir o rápido acesso a documentos relevantes com base na similaridade semântica, este tipo de mecanismo leva adicionalmente à melhoria contínua do processo de escrita dos documentos.

Os experimentos realizados com a ferramenta XMeaning mostraram a sua efetividade, contudo, se deram de forma limitada, no contexto do repositório das mesas de negociação do SUS e artigos científicos. Pretende-se como trabalhos futuros ampliar a sua utilização em outros repositórios, com documentos de estrutura heterogênea, ou seja, bases envolvendo documentos de naturezas distintas. Também pretende-se investigar outros algoritmos de extração de termos-chave, de forma a melhorar as métricas de performance na identificação do conjunto de termos (tags).

REFERÊNCIAS

MILLER, George A.; CHARLES, Walter G. Contextual correlates of semantic similarity. Language and cognitive processes, v. 6, n. 1, p. 1-28, 1991.

TAN, Ah-Hwee et al. Text mining: The state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. sn, 1999. p. 65-70.

BERRY, Michael W.; CASTELLANOS, Malu. Survey of text mining II. New York: Springer, 2008.

MINER, Gary. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

RAJA, Uzma et al. Text mining in healthcare. Applications and opportunities. J Healthc Inf Manag, v. 22, n. 3, p. 52-6, 2008.

KOH, Hian Chye et al. Data mining applications in healthcare. Journal of healthcare information management, v. 19, n. 2, p. 65, 2011.

AGGARWAL, Charu C.; ZHAI, ChengXiang (Ed.). Mining text data. Springer Science & Business Media, 2012.

TOMAR, Divya; AGARWAL, Sonali. A survey on Data Mining approaches for Healthcare. International

Journal of Bio-Science and Bio-Technology, v. 5, n. 5, p. 241-266, 2013.

HOTHO, Andreas; NÜRNBERGER, Andreas; PAAß, Gerhard. A brief survey of text mining. In: Ldv Forum. 2005. p. 19-62.

GANTZ, John; REINSEL, David. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, v. 2007, n. 2012, p. 1-16, 2012.

VALIPOUR, Mohammad Hadi et al. A brief survey of software architecture concepts and service oriented architecture. In: Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on. IEEE, 2009. p. 34-38.

MILANOVIC, Nikola. Contract-based web service composition framework with correctness guarantees. In: ISAS. 2005. p. 52-67.

CHRISTENSEN, Jason H. Using RESTful web-services and cloud computing to create next generation mobile applications. In: Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications. ACM, 2009. p. 627-634.

BELIGA, Slobodan; MEŠTROVIĆ, Ana; MARTINČIĆ-IPŠIĆ, Sanda. An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences, v. 39, n. 1, p. 1-20, 2015.

RODRIGUEZ-MURO, Mariano; KONTCHAKOV, Roman; ZAKHARYASCHEV, Michael. Ontology-based data access: Ontop of databases. In: International Semantic Web Conference. Springer, Berlin, Heidelberg, 2013. p. 558-573.

HASAN, Kazi Saidul; NG, Vincent. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010. p. 365-373.