Revista de Informática Aplicada, Volume 11, Número 2, 2015

07

Metodologia para Mineração de Dados em Fóruns do Moodle: um estudo de caso para Gestão Educacional

Leandro A. Silva[1]

Abstract. *The use of social networks to disseminate news and opinion monitoring has been as strategic decision-making for the most diverse areas of knowledge. However, due to the large volume of publications and interactions in social networks, the knowledge extraction is not a trivial task to be done by a human, and therefore, data mining techniques have been widely used for this purpose. This work uses forums of the learning platform Moodle, considered here in this work as a pseudo social network, and proposes a methodology that involves data clustering, text mining and data visualization techniques to discovery knowledge from Fórum publications. The methodology presented and the generated indicators can be easily applied to any other type of Moodle forum use as, for example, analysis of learning. In this approach, the methodology is used to monitor and to measure the quality of management educational.*

Keywords: *Educational data mining. Data clustering. Social network. Moodle.*

Resumo. *O uso de redes sociais para divulgação de notícias e monitoração de opinião tem sido estratégico como tomada de decisões para as mais diversas áreas do conhecimento. Contudo, devido ao grande volume de publicações e interações em redes sociais, a extração de conhecimento não é uma tarefa trivial para ser feita por um humano e, por isso, técnicas de mineração de dados vêm sendo largamente utilizadas para essa finalidade. Este trabalho utiliza fóruns de discussões da ferramenta de apoio de ensino Moodle, considerado neste trabalho como uma pseudo rede social, e propõe uma metodologia que envolve agrupamento de dados, processamento de textos e técnicas de visualização de dados para extrair conhecimento das publicações do Fórum. A metodologia apresentada e os indicadores gerados podem ser facilmente aplicados para qualquer tipo de uso do fórum do Moodle como, por exemplo, em análise de aprendizagem. Neste trabalho será estudado um caso para o monitoramento e medida de qualidade da gestão educacional.*

Palavras-chave: *Mineração de Dados Educacionais. Agrupamento de dados. Redes sociais. Moodle.*

1. Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie - São Paulo - Brasil prof.leandro.augusto@mackenzie.br

1. Introdução

Com o advento das redes sociais e a facilidade em seu acesso, o uso estratégico dessas plataformas tem sido feito por diferentes áreas de negócio. Casos típicos de uso das redes sociais são: empresas que desejam lançar um novo produto ou serviço e usam a rede social para medir o grau de aceitação dos seus clientes; políticos que desejam saber sobre as principais necessidades de seus eleitores; agentes policiais e de inteligência desejam descobrir mensagens que possam colocar em risco a sociedade; ou mesmo instituições de ensino que desejam se internacionalizar ou ampliar seus métodos de ensino e estão interessadas em saber o que pensam os seus educandos (ou futuros educandos); entre outros casos de aplicação. Com o amplo uso desse veículo de comunicação, o volume de dados gerado torna a análise e a interpretação de dados tarefas difíceis aos humanos e, então, a mineração de dados emerge como uma importante ferramenta para auxiliar a descoberta de conhecimento nas publicações das redes sociais, transformando dados não estruturados em informações estratégicas para uso em tomadas de decisões.

Ao mesmo tempo em que a rede social pode ser usada a favor das empresas, existem situações em que as mesmas desejam evitar a exposição nesse tipo de mídia. Por exemplo, quando se faz a publicação de um problema em uma rede social, o que poderia talvez ser resolvido de forma simples toma outras dimensões de solução, por vezes, difícil de ser gerenciada. Não se trata, entretanto, de uma publicação na rede social ter a exposição de sentimentos e, que as empresas estejam apenas interessadas em coletar opiniões para avaliar produtos, monitorar marcas e estudar comportamentos de usuários (AGARWAL et al., 2011). O fato é que em algumas situações algum descontentamento individual, que pode não ser a opinião do coletivo, acaba tomando proporções desnecessárias. E, ainda, deve-se considerar também a possibilidade da existência de notícias falsas e irresponsáveis que podem ser publicadas para denegrir uma pessoa ou marca.

Com o foco na educação superior, tema central deste artigo, as redes sociais podem ser usadas como ferramentas de uso prático ao ensino e à gestão educacional. No ensino, quando se considera que os estudantes estão frequentemente conectados nas redes sociais e poderiam usá-las para discutir temas específicos, complementar um tópico da aula ou usar como ambiente de discussão (SILVA et al., 2015).

Por outro lado, a Gestão de Instituições de Ensino Superior (IES) é uma tarefa difícil, geralmente executada por um Diretor e um ou mais Coordenadores de Curso. O grande desafio dos gestores em administrar uma Faculdade ou um Curso constitui-se de aspectos acadêmicos, como atualização de projetos pedagógicos, gestão de professores etc.; administrativos, através da supervisão das salas de aula, instalações físicas etc.; e muitos outros desafios institucionais ou até mesmo políticos (TACHIZAWA; ANDRADE, 1999; TANAKA; PESSONI, 2011). A responsabilidade da administração ainda pode interferir na qualidade do ensino como um todo ou, então, se tornar motivo do aluno evadir de um curso (BAGGI; LOPES, 2011).

Ampliando a discussão na gestão do ensino, que será o estudo de caso deste trabalho, quando algum problema gerencial acontece, muitas vezes os gestores acabam sendo as últimas pessoas a saber da situação, e isso eleva a dificuldade da solução. O ideal, nesse caso, é o gestor saber do problema no momento em que ele surge, ou próximo à ocorrência.

Pensando nisto, este artigo apresenta uma proposta que usa o recurso do Moodle (*Modular Object-Oriented Dynamic Learning Environment*) chamado Fórum, possibilitando aos alunos de uma IES, devidamente cadastrados nesse ambiente, a possibilidade de publicar dúvidas, discussões e até mesmo as ocorrências vivenciadas em sua rotina escolar (MOODLE, 2016). Por essas características do Moodle e, considerando também ser um ambiente que conecta um grupo alunos, professores e gestores para uma discussão coletiva, neste trabalho será adotada a definição de Recuero et al. (2009) que define o Moodle como uma *pseudo* rede social.

Para analisar as publicações do fórum do Moodle, independentemente do conteúdo ser uma dúvida de aluno ou uma ocorrência na rotina escolar, deve-se ressaltar que são dados de natureza textual e, portanto, é necessário considerar a utilização de um processo com técnicas de linguagem natural (PLN) para tornar as publicações passíveis de análise. Nesse processo, as publicações do fórum ou *corpus*, como esse tipo de dados é definido em PLN, são tratadas para a construção de uma coleção de palavras ou *BoW* (do inglês, *bag-of-words*). A partir desse dicionário formado pela *BoW* cada publicação é representada numericamente e o resultado é um conjunto de dados estruturado e pronto para análise (SILVA; PERES; BOSCARIOLI, 2015).

A partir do *corpus* representado, como o resultado final da representação é chamado, aplicam-se as técnicas de Mineração de Dados (MD) (SILVA, 2015). Como exemplo de aplicação da MD, as publicações podem ser agrupadas por similaridade de conteúdo. Esse processo de estruturar a publicação e realizar o agrupamento de dados com o algoritmo *k*-médias fez parte do estudo preliminar a esta pesquisa (SILVA et al., 2015). Aqui, a proposta é apresentar uma metodologia de descoberta de conhecimento para publicações do fórum do Moodle, com uso em qualquer contexto acadêmico. Adicionalmente a essa contribuição, apresenta-se também como parte deste trabalho uma estratégia para que o *corpus* representado e agrupado seja analisado com uso de nuvens de palavras (*word cloud*) e transformado em indicador de qualidade de um processo. Na análise da nuvem de palavras, o especialista do domínio (gestor ou professor) é envolvido no processo para fazer a categorização das postagens agrupadas em, por exemplo, burocracia, infraestrutura, avaliação, planejamento, comunicação etc. Ao resultado categorizado faz-se uso de análises exploratórias de dados como contagem e gráfico de frequência (Diagrama de Pareto) com o objetivo de transformar essas informações descobertas em indicadores de prioridades de solução e monitoramento de problemas descobertos (TRIOLA, 2008). A metodologia proposta neste trabalho pode ser aplicada em qualquer tipo de uso do fórum do Moodle como, por exemplo, para categorizar e monitorar dúvidas de tópicos de uma disciplina.

O trabalho está organizado como segue. No Capítulo 2 discute-se, de forma introdutória, o Moodle, onde assume-se aqui neste trabalho como sendo uma pseudo rede social. No Capítulo 3, apresenta-se uma introdução sobre Mineração de Dados e a descrição do algoritmo usado neste trabalho. A metodologia proposta neste trabalho está no Capítulo 4. No Capítulo 5 apresenta-se os resultados e análises dos experimentos realizados. Por fim, as conclusões são apresentadas.

2. Rede Social

Com a evolução tecnológica, as redes sociais físicas passaram para o âmbito eletrônico e hoje temos as mídias sociais também conhecidas como redes sociais, que é um conjunto de pessoas ou entidades interligadas uns aos outros através da Internet (RECUERO, 2009). Atualmente existem diversas redes sociais, cada uma abrangendo um nível específico da vida social, tais como: redes de relacionamentos, redes profissionais, redes comunitárias, redes políticas, redes educacionais. Alguns exemplos de redes sociais são:

- **Facebook:** é uma rede social de relacionamento onde o usuário cria seu perfil, adiciona amigos, publica fotos, vídeos e *links* que são compartilhados com todos os amigos.
- **Twitter**: é uma rede social de relacionamento onde o usuário cria um perfil e a partir disso pode acompanhar outros perfis de acordo com seu interesse; e também poderá ser acompanhado pelos usuários interessados.
- **Youtube**: é uma rede social destinada a vídeos onde o usuário cria seu perfil e atrela a ele seus vídeos favoritos e publicados. Todos esses vídeos ficam disponíveis em uma página chamada de canal.
- **LinkedIn:** é uma rede social profissional onde o usuário cria seu perfil com todos os dados profissionais e acadêmicos com o objetivo de se relacionar com outros profissionais. O uso dessa rede ainda permite indicar perfis a vagas de empregos.

As redes sociais possuem uma grande capacidade de geração de dados, principalmente pelo fato da quantidade de usuários que as acessam e nelas interagem. Entender as vantagens dessa nova forma de comunicação que possui um considerável alcance e utilizar todas essas informações geradas diariamente, a fim de gerar conhecimentos, poderão trazer grandes benefícios e até mesmo ganhos financeiros em diversas áreas do saber. Na educação ainda não existe um tipo de rede social e, portanto, pela razão do ambiente de auxílio ao aprendizado Moodle proporcionar conexões às pessoas há autores que consideram o ambiente como uma *pseudo* rede social, embora a rede não possa ser gerenciada pelos usuários (RECUERO, 2009). Portanto, a *pseudo* rede social Moodle permite a troca de informações entre os usuários, aumentando, assim, o conhecer de cada usuário através de suas iterações no ambiente como a troca de mensagens.

No Moodle existe o recurso chamado fórum, onde os alunos, participantes dessa *pseudo* rede social, publicam assuntos diversos acerca de um assunto de aula ou mesmo da Instituição de Ensino Superior

(IES). Por essa razão, o objetivo deste trabalho é verificar os principais assuntos discutidos na atividade fórum do Moodle de uma IES e, com isso, descobrir conhecimento que pode ajudar gestores ou professores em estratégias para a tomada de decisão.

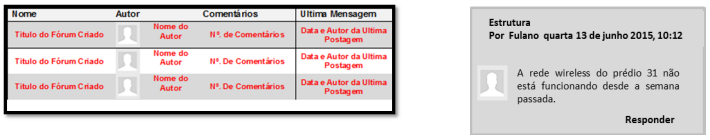
A extração de conhecimento no fórum do Moodle não é uma tarefa trivial para análise humana, pois as publicações são volumosas e se dão de forma textual e não estruturada. Para um processo de descoberta de conhecimento automático ou semi-automático com uso de Mineração de Dados, esse tipo de dados precisa ser transformado em um formato numérico. Na seção seguinte será explicado com mais detalhes como se faz essa transformação e a tarefa de mineração de dados que será aplicada neste estudo.

3. Mineração de Dados

Mineração de Dados (MD) é uma área de pesquisa multidisciplinar, envolvendo basicamente Banco de Dados, Estatística e Aprendizado de Máquina. A MD é parte principal de um processo que tem como entrada uma Base de Dados e como saída um Conhecimento (FAYYAD; PIATETSKI-SHAPIRO; SMYTH, 1996). Ela é divida em tarefas como predição, agrupamento e associação de dados que devem ser escolhidas de acordo com análises exploratórias, inicialmente feitas sobre os dados disponíveis para análise (HAN; KAMBER, 2006; TAN; STEINBACH; KUMAR, 2009; FACELI et al., 2011; SILVA, 2015). A MD tem sido amplamente utilizada em diferentes áreas, inclusive na educação, a qual é chamada como Mineração de Dados Educacionais ou MDE (SILVA; SILVA, 2014; ROMERO; VENTURA, 2010).

3.1 Mineração de Texto

Um fórum do Moodle é formado por um conjunto de mensagens estruturadas em nome da publicação, identificação do autor da publicação (foto e nome), quantidade de comentários sobre a publicação e a data do último comentário, como se pode ver no exemplo da Figura 1a. A publicação, por outro lado, é constituída pelo título, data de postagem, o texto da publicação e uma opção de resposta, como mostrado na Figura 1b.



a) exemplo de um conjunto de publicações. b) exemplo de uma publicação.

Figura 1 - Exemplo de estrutura de publicação do Fórum do Moodle.
Fonte: autor.

Para a aplicação de algoritmos de análises de dados, como a tarefa de agrupamento da Mineração de Dados (explicada a seguir), cada publicação em formato de texto (Figura 1b) deve ser transformada em uma representação numérica. O processo responsável por essa transformação está ilustrado na Figura 2 (SILVA; PERES; BOSCARIOLI, 2016).

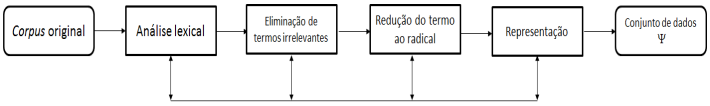


Figura 2 - Processo de construção
Fonte: Silva, Peres e Boscarioli, 2016.

O conjunto de postagem é chamado de *corpus* e é a entrada do processo de representação, cujas fases subsequentes a essa inicial ocorre da seguinte maneira:

- *Análise lexical*: a partir do caractere espaço, as palavras das publicações são separadas em termos ou *tokens*. Os caracteres especiais como vírgulas (“,”), pontos de exclamação (“!”) e de interrogação (“?”), entre outros, são removidos, assim como os números. Aqui também pode-se remover os marcadores de HTML ou XML presentes em alguns casos, como o das postagens do Moodle.
- *Eliminação de termos irrelevantes*: palavras como preposições, pronomes, artigos, advérbios que são comuns a ambos os vocabulários são removidas do processo. Essas palavras são também conhecidas como *stopwords*;
- *Redução do termo ao radical*: as palavras resultantes das etapas anteriores passam por uma normalização ortográfica para que sejam reduzidas ao radical. Esse processo

é importante, pois permite que palavras com o mesmo radical sejam consideradas como semelhantes. Esse processo também é conhecido como *stemming*.

Ao final dessa etapa do processo gera-se um vocabulário de palavras chamado de *bag-of-words* (*BoW*). O *corpus* representado é gerado a partir desse vocabulário. A representação desse *corpus* pode ser feita, por exemplo, usando um modelo chamado vetorial (GÖKER; DAVIES, 2009). O modelo vetorial busca representar o *corpus* em forma numérica, atribuindo pesos a cada termo (palavra) da *BoW* presentes na publicação, indicando, assim, uma importância. Uma das formas de se fazer a ponderação é combinando a frequência que um termo aparece em um *corpus* ou *tf* (do inglês, *term of frequency*) com a frequência invertida do documento ou *idf* (do inglês, *inverse document frequency*). A combinação de *tf* x *idf* define a importância (peso) de um termo dentro do *corpus* (GÖKER; DAVIES, 2009; SILVA; PERES; BOSCARIOLI, 2016).

Assim, depois de calculado o peso de cada termo do documento (representação), tem-se como resultado um conjunto de dados, que é exatamente a representação numérica do conjunto de publicações a ser apresentada ao algoritmo de análise de agrupamento.

3.2 Análise de Agrupamento

Dado um conjunto de dados, neste trabalho, um *corpus* estruturado, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_i,.. \mathbf{x}_n,]$, o *i*-ésimo exemplar (publicação no fórum) desse conjunto é descrito por atributos como sendo, $\mathbf{x}_i = [a_1, a_2,..., a_j,... a_m]$. A descoberta de grupos ou *clustering* trata-se de um modelo que procura encontrar exemplares $\mathbf{x}_i$ com atributos (termos) a_j semelhantes no conjunto de dados disponível para análise, **X**. A segmentação da base em grupos é feita a partir de medidas de similaridade. Em problemas que envolvem agrupamento não se tem disponível o atributo especial classe ou rótulo do exemplar. Por essa razão, dizemos que o aprendizado realizado pelos algoritmos de agrupamento é não-supervisionado (JAIN; DUBES, 1988; HAN; KAMBER, 2006; TAN; STEINBACH; KUMAR, 2009; JAIN, 2010; FACELI et al., 2011; WITTEN et al., 2001; SILVA, 2015).

O aprendizado não-supervisionado é de grande desafio, pois não se tem o objetivo que se deseja alcançar (classe ou rótulo), o que significa não conhecer o número de grupos da base de dados. E, ainda, os exemplares estão distribuídos em um espaço de dimensão elevada com diferentes formatos e separação. Esses aspectos ilustram a grande dificuldade para lidar com problemas de agrupamento de dados.

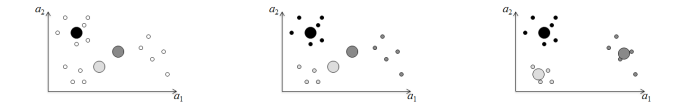
Em um contexto geral, o que o agrupamento faz é a descoberta de perfil. Mais especificamente na educação, a descoberta poderia ser útil para descobrir estilos de aprendizado dos alunos, padrões de erros, disciplinas de interesses comuns e muitas outras aplicações (SILVA; SILVA, 2014).

Os algoritmos de agrupamento também podem ser divididos em Aprendizado de Máquina e Inteligência Computacional. Algoritmos típicos de Aprendizado de Máquina são o Agrupamento Hierárquico, *k*-Médias (*k-Means*), Agrupamento Espacial baseado em Densidade ou DBScan (*Density Based Spatial Clustering of Applications with Noise*). Por outro lado, a abordagem baseada em Redes Neurais tem o Mapa Auto-Organizável ou SOM (*Self-Organizing Map*) (JAIN; DUBES, 1988; HAN; KAMBER, 2006; TAN; STEINBACH; KUMAR, 2009; HAYKIN, 2009; JAIN, 2010; FACELI et al., 2011; WITTEN et al., 2011; SILVA; PERES; BOSCARIOLI, 2016). A seguir, será apresentado em detalhes o *k*-Médias, algoritmo a ser utilizado neste estudo.

3.2.1 O algoritmo k-Médias

O *k*-Médias ou, em inglês, *k*-Means, é o principal algoritmo de agrupamento particional (*Partitional Clustering*) (SILVA, 2015). O objetivo do algoritmo é encontrar particionamentos nos exemplares do conjunto de dados dentro de *k* grupos disjuntos, sendo *k* um valor inteiro maior que um. O número de grupos deve ser dado como um parâmetro de entrada do algoritmo, como também a medida de distância - geralmente a Euclidiana. Com os parâmetros de entrada definidos é feita uma escolha aleatória dos *k* centroides de grupos. A partir desses centroides, o algoritmo iterativo gera as partições na base, fornecendo como resultado final a base de dados agrupada. A Figura 3 ilustra um exemplo da operação do algoritmo *k*-Médias. Na Figura 3a os exemplares são representados por dois atributos descritivos (a_1 e a_2), constituindo o conjunto de dados (círculos pequenos e sem preenchimento); nessa figura ainda apresenta-se os centroides (círculos

grandes e com preenchimento), cujas posições iniciais serão assumidas por escolhas aleatórias. Após parametrizações iniciais, cada exemplar é comparado com os centroides através de uma medida de distância, sendo que a menor distância indica uma associação do exemplar ao centroide, como mostra a Figura 3b, em que as associações são representadas pela atribuição de cores dos exemplares às do centroide mais próximo. Por fim, Figura 3c, o centroide é atualizado, a partir da média dos atributos dos exemplares nele agrupado. O Algoritmo 1 descreve todos os passos desse processo de agrupamento de dados.



a) Inicialização dos centroides b) Associação aos centroides mais próximos c) Atualização dos centroides

Figura 3 - Exemplo de operação do algoritmo k-Médias
Fonte: autor.

O algoritmo *k*-Médias explicado nesta seção espera que os exemplares parem de fazer troca de centroides em alguma iteração, o que pode não acontecer. Por essa razão, pode-se considerar uma função objetivo para que seja minimizada (ou maximizada) até um valor aceitável como critério de parada do algoritmo ou que ao menos se defina um número máximo de iterações para que o algoritmo se estabilize (TAN et al., 2011).

Algoritmo 1 - Algoritmo k-Médias

Parâmetros de Entrada:
- **X**: um conjunto de dados não rotulado;
- *k*: número de grupos que se deseja encontrar;
- *d*: métrica de distância;

Parâmetro de Saída:
- base de dados agrupada

1. escolha aleatoriamente *k* distintos valores para centroides de grupos
2. repita
3. para cada (exemplar de **X**) **faça**
4. calcule a distância *d* entre o exemplar escolhido e os centroides
fim para
5. agrupe o exemplar ao centroide mais próximo
6. recalcule o centroide para cada grupo
até *não haver mais alteração no agrupamento dos exemplares aos centroides*

Em problemas de agrupamento não há metodologia para validação de resultados. No entanto, ainda assim é possível medir desempenho. Em agrupamento, a medida de desempenho é chamada de índice de validação de agrupamento (em inglês, *cluster validity indices*) e os métodos se dividem em interno, relativo e externo (JAIN; DUBES, 1988). Os índices internos baseiam-se na estrutura gerada pelo algoritmo de agrupamento. Portanto, a grande maioria dos quantificadores mensuram a coesão e o isolamento dos grupos. Exemplos de índices internos são Coeficiente de Silhoueta, Índice de Dunn, Índice de Davies-Bouldin entre outros (SILVA; PERES; BOSCARIOLI, 2016).

4. Metodologia Experimental

Para a realização dos experimentos utilizou-se como base de dados as postagens publicadas em um fórum do Moodle de uma IES chamado de críticas e sugestões durante dois anos, totalizando 1000 publicações. A ferramenta utilizada nos experimentos foi a RapidMiner (2016), que possui diversos tipos de recursos para Mineração de Dados e análise de texto. A metodologia proposta no trabalho está ilustrada na Figura 4.

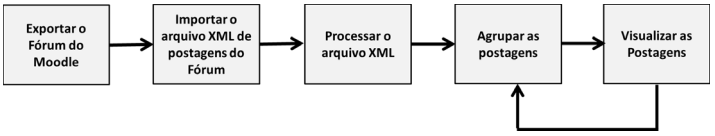


Figura 4 - Metodologia proposta neste trabalho
Fonte: dados da pesquisa.

Explicando cada fase do processo metodológico (Figura 4), os dados foram exportados do Fórum do Moodle em formato XML. A estrutura do arquivo contém uma série de informações além das postagens, por exemplo, autor da publicação, data, dia etc. Então, no processo de importação é preciso interpretar o arquivo para extrair apenas os dados de interesse, que neste trabalho estão identificados por SUBJECT e MESSAGE, veja na Figura 5 as linhas sombreadas. Embora o MESSAGE seja o texto que contém a postagem do aluno e tenha o conteúdo que será usado na análise efetivamente, o SUBJECT é importante para auxiliar na rotulação das publicações agrupadas.

Após a importação dos campos desejados do arquivo XML, o texto do MESSAGE precisa ser processado. Esse texto, como apresentado antes na Subseção 3.1, é chamado de *corpus* e precisa passar pelo processo da Figura 1 para que seja representado por um conjunto

de dados estruturado (em matriz), sendo cada linha uma publicação, cada coluna uma palavra e o valor a importância de cada palavra na publicação.

Com a representação por uma matriz de números o algoritmo *k*-Médias é aplicado. Como explicado antes na Subseção 3.1, o parâmetro *k* indica a quantidade de grupos a ser descoberto. Na prática, não se sabe qual é o valor ideal desse parâmetro. Então, na metodologia deste trabalho, o agrupamento será feito usando diferentes valores de *k*, iniciando em 2 e avaliando o desempenho pela soma quadrática, que mensura a coesão dos exemplares em um grupo (JAIN; MURTY; FLYNN, 1999). Além da análise quantitativa da coesão por meio da soma quadrática, cada grupo será analisado qualitativamente com uso da técnica de visualização de dados conhecida como nuvens de palavras (*world cloud*). Nessa técnica de visualização, considera-se as 10 palavras com maior frequência dentro de cada grupo. A combinação de análise qualitativa e quantitativa é uma estratégia que se mostrou importante neste trabalho, pois, embora o índice de soma quadrática seja uma medida objetiva, sendo o menor valor o agrupamento ideal, a nuvem de palavras permite saber se os assuntos gerais estão sendo especificados em detalhes. Por exemplo, um tipo de postagem para comunicação de aviso deveria ser assim considerado, de maneira genérica, ao invés de detalhar se refere a bolsas de estudo, eventos ou alguma outra atividade. Por isso, nessa etapa do processo, a de visualização, há uma realimentação para a etapa do agrupamento (veja Figura 4).

```
<POSTS>
 <POST>
  <ID>42303</ID>
  <PARENT>42271</PARENT>
  <USERID>90656</USERID>
  <CREATED>1276785712</CREATED>
  <MODIFIED>1276785712</MODIFIED>
  <MAILED>1</MAILED>
  <SUBJECT>Re: Prova Unificada</SUBJECT>
  <MESSAGE>Seria interessante ter o gabarito e respostas. </MESSAGE>
  <FORMAT>1</FORMAT>
  <ATTACHMENT></ATTACHMENT>
  <TOTALSCORE>0</TOTALSCORE>
  <MAILNOW>0</MAILNOW>
 </POST>

  <POST>
    ....
    </POST>
<POSTS>
```

Figura 5 - Exemplo da estrutura XML e exemplo de uma publicação (postagem).
Fonte: dados da pesquisa.

Finalmente os resultados de agrupamento serão quantificados em termos de grupos, postagens e enfim categorizados com o assunto central de cada grupo. O resultado pode, por exemplo, ser graficamente apresentado em frequência ordenada pelo maior valor, ou seja, um diagrama de Pareto (TRIOLA, 2008) para acompanhar o melhoramento, neste estudo, das críticas e sugestões da Unidade Universitária (UU) em análise.

5. Resultados Experimentais

O experimento iniciou-se com seleção das 1.000 publicações no fórum do Moodle. Através da ferramenta Rapid Miner realizou-se inicialmente alguns passos da preparação dos dados, como: o *tokenize* que separa cada palavra de um *post* em um termo e onde se fez o tratamento para retirar as marcações XML e a extração de *stopwords* que elimina termos como preposições, artigos e todos elementos gramaticais comuns a todas publicações. Após análise desse primeiro processo, notou-se a necessidade de inclusão da etapa de *stemming*, que extrai de cada termo o seu radical. Isso se deu pelas diferentes formas de tratamento das publicações, como, por exemplo, em relação ao coordenador como “prof.” ou “professor”. A importância dessas operações no processo de preparação de documentos relaciona-se também com o número de termos existentes, que, por sua vez, estão vinculados diretamente com a dimensão dos valores ponderados (*tf* x *idf*).

Na tarefa de Mineração de Dados, embora o algoritmo de descobrir grupos *k*-Médias seja de simples parametrização, como discutido anteriormente, a tarefa de agrupamento de dados não é trivial. A escolha do valor de grupos, *k*, passa a ser um problema quando não se sabe quantos grupos existem no conjunto de dados. Na prática, o que se faz é variar o valor de grupos, iniciando *k* com valor maior ou igual a 2 e, a cada variação, avalia-se a qualidade do grupo através de um índice. A Figura 6 mostra o resultado da investigação do número de grupos. Nota-se que, a partir de $k \geq 5$, a soma quadrática da distância entre os exemplares e a média do grupo começa a reduzir lentamente, sugerindo uma estabilização do valor de *k*. No entanto, é importante ressaltar que a decisão por oito grupos ($k = 8$), além de apresentar o menor resultado, que é o valor ótimo dessa análise e, por isso, já é suficiente como critério, a análise de agrupamento com a interpretação da nuvem de palavras possibilita concluir

que, para valores maiores a 8 haveria particionamentos das publicações em minigrupos, os quais resultariam em um volume pequeno de até duas publicações. É importante lembrar que a quantidade de publicações em cada grupo, e o relacionamento entre grupos são fatores importantes para a geração de conhecimento.

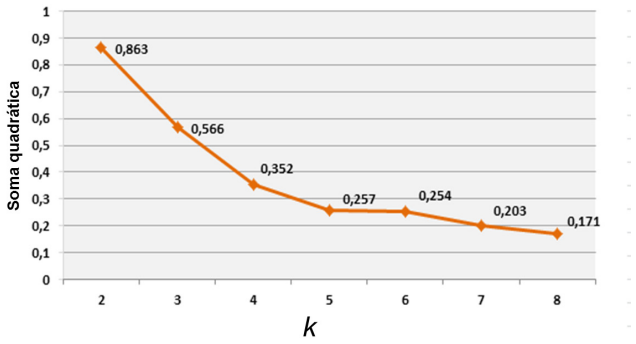


Figura 6 - Gráfico que apresenta a variação do índice de qualidade de agrupamento, soma quadrática, para os valores de *k* analisados.
Fonte: dados da pesquisa.

Para melhor analisar os resultados, as postagens mapeadas em cada grupo foram analisadas e visualizadas por nuvem de palavras (*word cloud*) e, a partir de então, foram criadas categorias para cada grupo de publicações. Por exemplo, a Figura 7 apresenta as dez palavras mais frequentes nas publicações de um dos agrupamentos descobertos pelo *k*-Médias. Por essa nuvem pode-se concluir que as publicações são manifestações dos alunos sobre o critério de notas de uma avaliação da Unidade Universitária (UU) que mensura a retenção de conhecimento do aluno semestralmente, deixando de ser uma nota de participação e passando a ser uma componente da média final. Ou seja, o agrupamento com essas mensagens será categorizado como "Prova unificada". E análise semelhante foi feita com os demais grupos descobertos.

O resultado dos agrupamentos descobertos está apresentado na Tabela 1. Nessa tabela estão o número de postagens em cada grupo, a porcentagem relativa e também a categoria atribuída. Ou seja, o grupo 1 tem o tema "Comunicado" referente a avisos da Faculdade sobre as parcerias estabelecidas com empresas, o que é um tipo de grupo com mensagens positivas, representando 11,5% de publicações. O grupo 2 tem mensagens referentes a falta de "Organização" da unidade e quantidade de "Burocracias", portanto, representando 29,2% das mensagens com manifestações sobre problemas na UU. No grupo 3 há mensagens de comunicado referentes a assuntos gerais dos cursos, bolsas de pesquisa, extensão e intercâmbio, ou seja, 10,8% de mensagens positivas com "Comunicado" aos alunos. No grupo 4 foram publicadas mensagens de falta de "Planejamento" da UU com algumas ações, portanto, 6,8% de mensagens de manifestação dos alunos. Agora, no grupo 5 há mensagens referentes à demora da coordenação em apresentar "Respostas" às dúvidas de alunos, sendo assim, 11,9% de mensagens de problemas. O grupo 7 é o exemplo apresentado antes sobre o problema com a "Prova Unificada", representando 5,6% de publicações. Finalmente, no grupo 8, há manifestações sobre problemas de rede sem fio, tomadas, cadeiras com defeito e vários outros problemas de "Infraestrutura", totalizando 5,3% de mensagens sobre problemas.

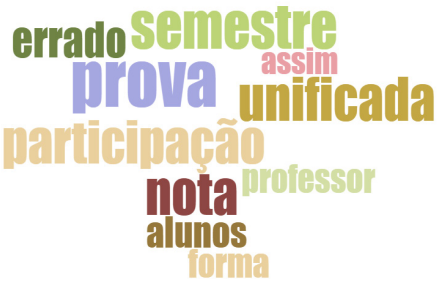


Figura 7 - Nuvem de palavras para um dos grupos gerados pelo k-Médias.
Fonte: dados da pesquisa.

A partir dos resultados dos grupos, Tabela 1, é possível se ter ideia do volume de postagens e, consequentemente, pode-se obter conhecimento às criticas, sugestões, ideias, pensamentos, impressões e constatações das principais publicações dos alunos. De maneira resumida, 77,7% das mensagens são referentes a manifestações dos alunos a problemas da UU em análise.

Tabela 1 - Resultados do Agrupamento.

Grupos	# Postagens	% Postagens	Categoria
1	115	11,5	Parcerias
2	292	29,2	Organização
3	108	10,8	Comunicação
4	68	6,8	Planejamento
5	119	11,9	Respostas
6	189	18,9	Avaliação
7	56	5,6	Prova unificada
8	53	5,3	Infraestrutura

Fonte: dados da pesquisa.

Para efeitos de monitoração do processo, gerou-se um Diagrama de Pareto, Figura 8, usando o número de postagens de publicação e a categoria de mensagens. Portanto, a UU deve tomar algumas ações com base nessa descoberta às postagens no Fórum do Moodle e, o diagrama de Pareto pode ser usado para acompanhar o efeito das ações realizadas.

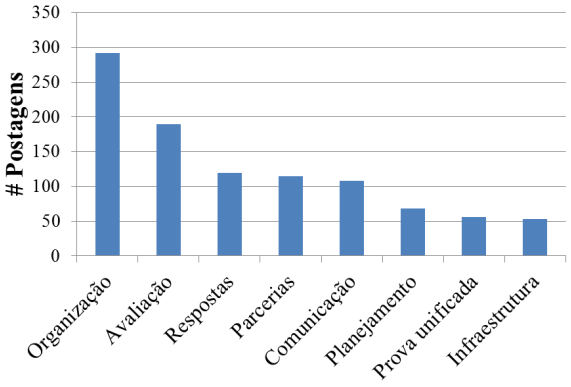


Figura 8 - Diagrama de Pareto das postagens descobertas no Fórum do Moodle.
Fonte: dados da pesquisa.

A metodologia apresentada neste trabalho pode também ser usada sobre postagens relacionadas a assuntos específicos de aula. Assim, as descobertas, a partir da análise dos grupos, envolveriam, por exemplo, os temas de maior repercussão ou de dúvidas. Ou seja, pode-se descobrir os assuntos mais compreendidos pelos alunos e os assuntos que mais geraram dúvidas.

6. Conclusão

Empresas, governos, instituições das mais diversas, incluindo de Ensino Superior, aplicam técnicas de Mineração de Dados para extrair informações de redes sociais que lhes podem ser úteis como informações estratégicas para a tomada de decisão. A Mineração de Dados, além de permitir revelar conhecimentos antes desconhecidos, também permite que se tenha redução no tempo dispensado para obtenção de informações, pois, se comparado às técnicas tradicionais de pesquisa e obtenção de opiniões, a Mineração de Dados nas redes sociais é mais rápida, visto que nos métodos tradicionais há a necessidade de criação de formulário de pesquisa, distribuição, preenchimento por parte dos pesquisados, computação dos dados obtidos e, por fim, a análise. Ou seja, todas essas etapas requerem muito tempo, o que não ocorre com os dados nas redes sociais. Vale ressaltar que nas pesquisas convencionais muitas vezes pode não se abranger os assuntos de maior interesse pelos alunos como, por exemplo: em uma pesquisa de satisfação a alunos, se o pesquisador não souber de antemão quais são os pontos mais preocupantes, então ele não poderá o incluir na pesquisa e essa informação será perdida. Por outro lado, na mineração de dados em redes sociais, os alunos poderão se expressar da maneira que melhor lhe convier sobre os diversos assuntos, expor seu ponto de vista e, após a mineração de dados, o conhecimento obtido ilustrará cada detalhe que seria omitido em formulário pré-fixados de pesquisa de opinião.

Neste trabalho, a aplicação da metodologia de Mineração de Dados em publicações no Moodle se tornou um processo importante por permitir a descoberta de grupos e, consequentemente, gerar índices que podem ser monitorados pelos gestores com o intuito de melhorar a qualidade da Gestão de uma Instituição de Ensino Superior. Como trabalho futuro pretende-se validar a metodologia com outros dados educacionais.

Referências

AGARWAL, A.; XIE, Boyi; VOVSA, I.; PASSONNEAU, R. Sentiment Analysis of Twitter Data. Columbia University. In: WORKSHOP ON LANGUAGES IN SOCIAL MEDIA. *Proceedings...* Association for Computational Linguistics, pp. 30-38, 2011.

BAGGI, C. A. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação*, Campinas; Sorocaba, SP, v. 16, n. 2, p. 355-374, 2011.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. Rio de Janeiro: LCT, 2011.

FAYYAD, U.; PIATETSKI-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining,* California. p. 1-34, 1996.

GÖKER, A.; DAVIES, J. *Information retrieval:* searching in the 21st century. Chichester, U.K.: John Wiley & Sons, Inc., p.6-7, 2009.

HAN, J.; KAMBER; M. *Data Mining Concepts and Tecniques.* 3. ed. San Francisco: Elsevier; Morgan Kaufmann, 2006.

JAIN, A. K.,; DUBES, R. C. *Algorithms for clustering data.* New Jersey: Prentice-Hall, 1988.

JAIN, A.K.; MURTY, M.N.; FLYNN, P.J. Data clustering: a review. ACM *Comput Surv*, New York, v. 31, n. 3, 264–323, 1999.

MOODLE. Modular Object-Oriented Dynamic Learning Environment. 2016. Disponível em: <https://moodle.org/>. Acesso em: 04 out. 2016.

RAPID MINER. Rapid Miner. 2016. Disponível em: <https://rapidminer.com/>. Acesso em: 04 out. 2016.

RECUERO, R. *Redes Sociais na Internet.* São Paulo: Meridional, 2009.

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, v. 40, n. 6, 601-618, 2010.

SILVA, L. A; PERES, S. M.; BOSCARIOLI, C. Introdução à Mineração de dados com aplicações em R. Rio de Janeiro: Editora Elsevier, 2016.

SILVA, L. A; TRINDADE, D.; DE PAULA, C.; PINTO, S. N. Mineração de Dados em publicações de Fóruns de Discussões do Moodle como geração de Indicadores para aprimoramento da Gestão Educacional. In: WORKSHOP DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. *Anais...* Trilha Mineração de Dados Educacionais, 1084-1093, 2015.

SILVA, L. A. *Mineração de dados:* uma abordagem introdutória e ilustrada. São Paulo: Editora Mackenzie (Coleção conexão inicial, v. 11), 2015.

SILVA, L. A.; SILVA, L. Fundamentos de Mineração de Dados Educacionais. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, *Anais...* v. 3, n. 1, 2014.

TACHIZAWA, T.; DE ANDRADE, R. O. B. *Gestão de instituições de ensino*. Rio de Janeiro: FGV Editora, 1999.

TAN, P.N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining:* mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

TANAKA, V. R. D. S.; PESSONI, L. M. D. L. A gestão do ensino superior: o gestor e seu papel. In: SEMINÁRIO SOBRE DOCÊNCIA UNIVERSITÁRIA, *Anais...* v.1, n. 1, 2011.

TRIOLA, M. F. *Introdução à Estatística.* 10. ed. Rio de Janeiro: Editora LTC, 2008.

WITTEN, I. H.; FRANK, E.; MARK, A.; HALL. *Data Mining:* Practical machine learning tools and techniques. 2. ed. San Francisco: Morgan Kaufmann, 2011.