SaOC: Sistema de Apoio à Otimização Curricular com uso de Mineração de Texto

Dhouglas Z. Martins, César A. D. Alves, Abilio R. Coelho, Marili M. S. Vieira, Nizam Omar, Leandro A. Silva

Faculdade de Computação e Informática – Universidade Presbiterana Mackenzie (UPM) – São Paulo, SP - Brazil

{dhouglas.zeraibe, cesar.diez}@mackenzista.com.br, {abilio.coelho, marili.vieira, omar, prof.leandro.augusto}@mackenzie.br

Abstract. *This research article describes the implementation of the Application Process of a Curricular Optimization Support System (SaOC), used as a proposal to optimize the common use of Curricular Components (CC) in a University, taking as an informative basis the comparison of similarity between the name, number of credits and format of classes (on-campus, distance learning and laboratory). Involving text mining techniques to apply similarity measures such as cosine and euclidean distance.*

Resumo. *Este artigo descreve a proposta de um processo e do desenvolvimento de um Sistema de Apoio a Otimização Curricular (SaOC), utilizando técnicas de Mineração de Dados Textuais para otimizar o compartilhamento de disciplinas, referenciadas aqui como Componentes Curriculares, de uma Universidade, tomando como base informativa a medida de similaridade entre o nome, número de créditos e formato de aulas (presenciais, à distância e laboratório). Adicionalmente, um estudo comparativo entre as medidas de similaridade cosseno e euclidiana foi realizado com o objetivo de verificar a eficiência de uso no sistema e os resultados analisados por meio de medidas de precisão e cobertura.*

1. Introdução

A Universidade consiste em uma Instituição de Ensino Superior pluridisciplinar formada por um conjunto de Faculdades em diferentes áreas do conhecimento. Um importante papel da Universidade é a socialização do conhecimento, que pode ser conseguido a partir da interação entre as Faculdades, compartilhando disciplinas que são comuns aos diferentes cursos. Isso deve-se principalmente ao existir no meio acadêmico atual diversas disciplinas com tópicos em comum e independente da área do conhecimento [Oliveira *et al.,* 2015].

No entanto, devido a diferentes aspectos como, por exemplo, políticos ou de especificidades de cursos, algumas disciplinas, aqui chamadas de Componentes Curriculares (CC), acabam não sendo oferecidas de maneira compartilhada. Apenas a título de exemplificação, tendo como base Universidades com curso de Matemática, denota-se que nem sempre o mesmo é responsável pelas CC relacionadas a esse curso em outras escolas, fazendo com que possam existir as mesmas CC, às vezes com nomes diferentes, mas com conteúdo semelhante. Ou seja, disciplinas comuns entre cursos,

DOI: 10.5753/cbie.wcbie.2017.634

caracterizadas por terem semelhança no nome, conteúdo curricular e bibliográfico, são vistas como sendo diferentes em uma mesma Universidade [Oliveira *et. al.,* 2015].

O problema abordado nesta pesquisa consiste em como identificar CC com semelhança no nome, na composição de crédito e no formato de aula (presenciais, laboratório ou EAD). Portanto, o objetivo deste artigo é propor e implementar um sistema que otimize o uso comum de CC, tomando como base informativa a similaridade entre o nome das disciplinas de uma Universidade Particular do Estado de São Paulo. Como objetivo específico, será feito um estudo comparativo das medidas de similaridade Cosseno e Euclidiana, e para verificar a eficiência e eficácia no processo de descobrir nomes de disciplinas similares.

O desenvolvimento do projeto se dará por meio de uso de técnicas de Mineração de Texto [Silva *et. al.,* 2016] para a representação de títulos de 1777 disciplinas distribuídas em 29 cursos de uma Universidade Particular. Após representação, será feito o uso de medidas de similaridade, comparando a distância Cosseno e Euclidiana, verificando o desempenho de cada uma delas com uso da medida de precisão por recuperação. Por fim, ao escolher a medida que apresenta melhor desempenho, apresenta-se a CC similares em pares, permitindo ainda visualizar outras informações como carga horária, formato das disciplinas e a faculdade que as oferta.

Além da introdução que tem como finalidade a contextualização da proposta de trabalho, o artigo está separado em outras cinco sessões as quais referem-se, respectivamente sobre os trabalhos correlatos à proposta de trabalho, a discussão da técnica de mineração de dados a ser aplicada, envolvendo assim a preparação dos dados textuais, a representação textual em valores numéricos e as medidas de similaridade que serão empregadas na análise. A quarta sessão está focada na metodologia utilizada, seguido então dos resultados levantados e, por fim, a sexta e última sessão contendo as conclusões obtidas.

2. Trabalhos Correlatos

Partindo de trabalhos que abordam problemas relacionados a análise de dados educacional ou acadêmica, IDA (2009) aplicou técnicas de *text mining* em dados educacionais, mais especificamente dos currículos base dos departamentos de Universidades, com a finalidade de criar um sistema de análise curricular, baseando-se no conceito de agrupamento e análise dos componentes extraídos (título, área de estudo etc.). O resultado da pesquisa possibilitou aprofundar a compreensão global sobre a informação utilizada nos currículo, conduzindo à descoberta de conhecimentos como termos técnicos dos conteúdos programáticos e o cálculo de graus de semelhança entre os conteúdos.

Ainda envolvido no contexto educacional, KAWINTIRANON *et. al*. (2016) propuseram uma abordagem baseada também em mineração de texto, a qual tem como objetivo avaliar o conteúdo programático de um curso; no artigo, é apresentada uma análise do conteúdo através da extração das ementas, no aspecto avaliativo do curso. A abordagem emprega técnicas de mineração de texto que extrai palavras-chave, utilizando a frequência de incidência das mesmas ao decorrer dos documentos. A análise é baseada em palavras-chave dos conteúdos programáticos de cursos, fazendo também uso da comparação para verificar a similaridade entre ementas, tomando como referência a geração de um valor atribuído através de um cálculo de *score* de similaridade entre os

documentos envolvidos, tendo como principal objetivo a proposição de um método alternativo de análise curricular para compreensão e avaliação do conteúdo das ementas em diversos cursos distintos e suas similaridades.

3. Mineração de Dados Textuais

Em Pezzine (2016), a mineração de dados textuais é retratada com base nos conceitos de descoberta de conhecimentos em bases de Dados e Mineração de Dados, trilhando o caminho para o descobrimento de conhecimentos correlatos a bases de dados textuais. Sendo assim, em conglomerados textuais, cada item pertencente a um conjunto de dados é tratado como uma espécie de 'documento', onde cada documento de corpus pode assumir diferentes características como, por exemplo, comprimento do texto, conteúdo etc. Nesse estudo, o *corpus* será utilizado de modo que contenha os documentos pertinentes à análise do componente curricular, partindo do preceito definido segundo Silva *et. al.,* (2016), um *corpus* ser apresentado como um grupo de documentos no formato de texto, portanto, de conteúdo não estruturado.

3.1. Preparação de dados textuais

A preparação do corpus (dado não estruturado) é constituída por um processo que tem início através da preparação dos dados e, segundo LAHITANI e colaboradores (2016), a primeira parte desse processo é feita com a análise lexical, gerando como resultado uma lista de *tokens*, portanto é uma etapa também conhecida como tokenização. Essa etapa é feita com a separação de cada palavra do documento e a eliminação de caracteres de pontuação como vírgula, ponto-e-vírgula, exclamação, interrogação etc.

Em Silva *et. al.,* (2016), a segunda etapa do processo consiste na remoção de *topwords* que se refere às palavras que não agregam informações discriminantes à tarefa de classificação de texto. Sendo assim, deve-se remover essas palavras dos documentos de texto. Alguns exemplos de *stopwords* são: a, sobre, depois, um, e, qualquer, de etc.

A redução dos termos ao radical é realizada de modo a permitir uma padronização semântica dos termos, possibilitando que palavras com o mesmo radical sejam tratadas de maneia conjunta [BADAWY, 2016].

3.2. Representação textual em valores numéricos

Nesta etapa, o *corpus* deve estar representado para ter sua representação vetorial gerada. Para realizar a representação textual em valores numéricos, comumente faz-se uso da combinação da frequência dos termos com o inverso da frequência nos documentos, a qual é uma medida que tem como objetivo definir a importância de cada palavra dentro de um documento.

A frequência dos termos (*tf*), indica diretamente a quantidade de vezes que um termo é encontrado em um documento. Em contrapartida, o inverso da frequência nos documentos (*idf*) é a raridade de cada palavra no conjunto de documentos. Portanto, a combinação do *tf* com *idf* representa um valor que pondera a frequência com a raridade de cada palavra no corpus.

3.3. Medidas de Similaridade

A partir do *corpus* representado em termos numéricos, pode-se então realizar manipulações analíticas sobre os textos como, por exemplo, aplicar técnicas de medidas

de similaridade para que seja evidenciada a correlação entre os documentos utilizados. Segundo GOMAA (2013) uma métrica mede a similaridade ou dissimilaridade (distância) entre dois objetos quaisquer para a comparação de sequências entre os caracteres dos documentos.

A distância cosseno, segundo LAHITANI (2016), mede a orientação do ângulo cosseno entre dois objetos (documentos). Em situações em que os objetos têm a mesma orientação, a similaridade cosseno será 1 e, caso contrário, objetos a 90º, a similaridade é 0. Portanto, independente da orientação, a distância cosseno sempre resultará em um valor positivo e normalizado entre 0 e 1. A equação da distância cosseno é assim definida:

$$Cos_{(\mathbf{A},\mathbf{B})} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \quad (1)$$

Sendo *a* e *b* componentes dos objetos sob análise, *n* o número de tokens e **A** e **B**, respectivamente.

Outro exemplo de métrica é conhecida como distância euclidiana, onde, segundo ZHANG (2016), é uma medida que descreve as dissimilaridades entre dois objetos. Sendo assim, objetos semelhantes terá o valor resultante igual a 0 e, caso contrário, quando mais dissimilar for o objeto, maior será o valor da distância Euclidiana. A equação dessa medida para os mesmos dois objetos **A** e **B** apresentado acima é:

$$Euclidiana_{(\mathbf{A},\mathbf{B})} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (2)$$

Neste trabalho um estudo comparativo destas duas medidas de similaridade será realizado, como será detalhado na seção seguinte.

4. Metodologia

Com base nos conceitos de Mineração de Dados Textuais explicados acima, este trabalho propõe o processo descrito na Figura 1. Este, inicia-se com a coleta de 1777 Componentes Curriculares (CC) distribuídas entre 12 Faculdades de uma Universidade Particular (seleção). Após isso, uma série de processamento é realizada para a transformação dos nomes das CC em dados estruturados para análise (pré-processamento, tokenização, *stopwords* e *stemming*), permitindo a representação numérica desses nomes (representação). Com as CC representadas, aplica-se uma medida de similaridade, distância Cosseno ou Euclidiana (similaridade), a fim de fazer descobertas sobre as disciplinas com nomes parecidos; e, por fim, apresenta-se as CC similares, permitindo ainda que se visualize outras informações das mesmas como carga horária, formato de disciplinas e Faculdade que a oferece (visualização).

O processo implementado resultou no Sistema de apoio a Otimização de Componentes Curriculares – SaOC, cuja interface está ilustrada na Figura 2. Nessa, note que a comparação é feita aos pares definindo o Conjunto 1 e 2 em que se caracteriza pelas variáveis campi, escola, curso e componente curricular. Como parâmetro para a apresentação das CC similares, o usuário pode definir um faixa de valores entre 30 e 100, sendo valores normalizados que traduzem as medidas de similaridade (para o caso da

distância Cosseno esse valor é simplesmente a multiplicação por 100). Como resposta, o sistema apresenta as disciplinas aos pares e com o nível de similaridade, possibilitando ainda que se visualize outras informações relacionadas as CC (carga horária, forma de oferecimento etc. como será explicado a seguir).

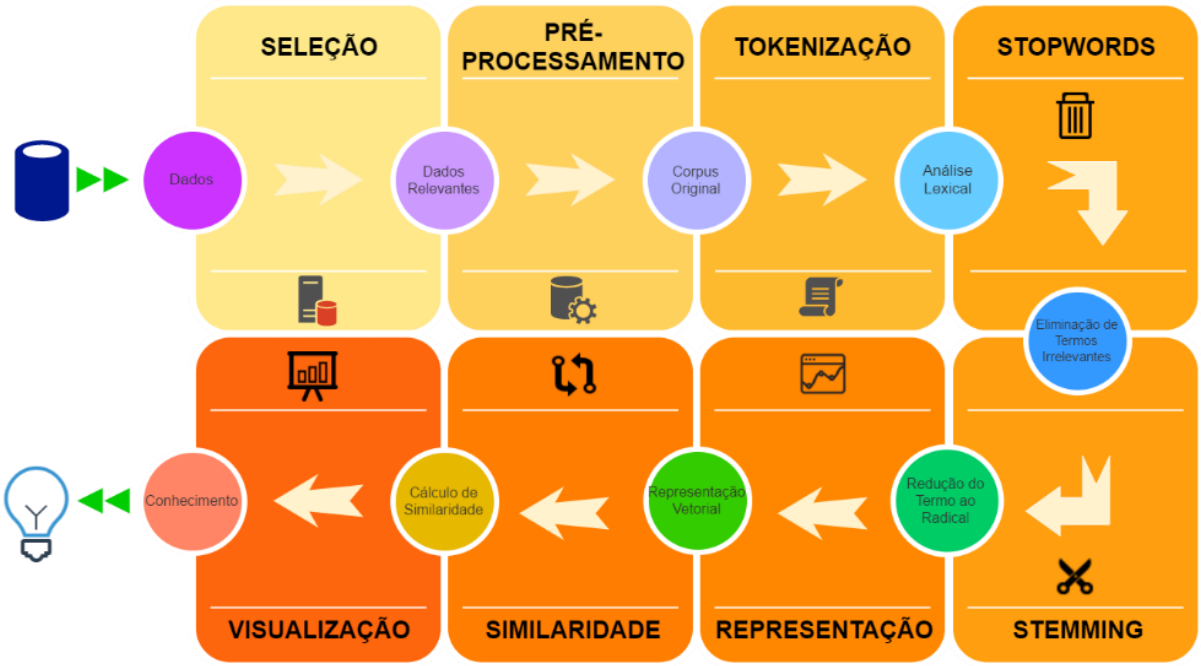

Figura 1 - Processo de proposto no projeto

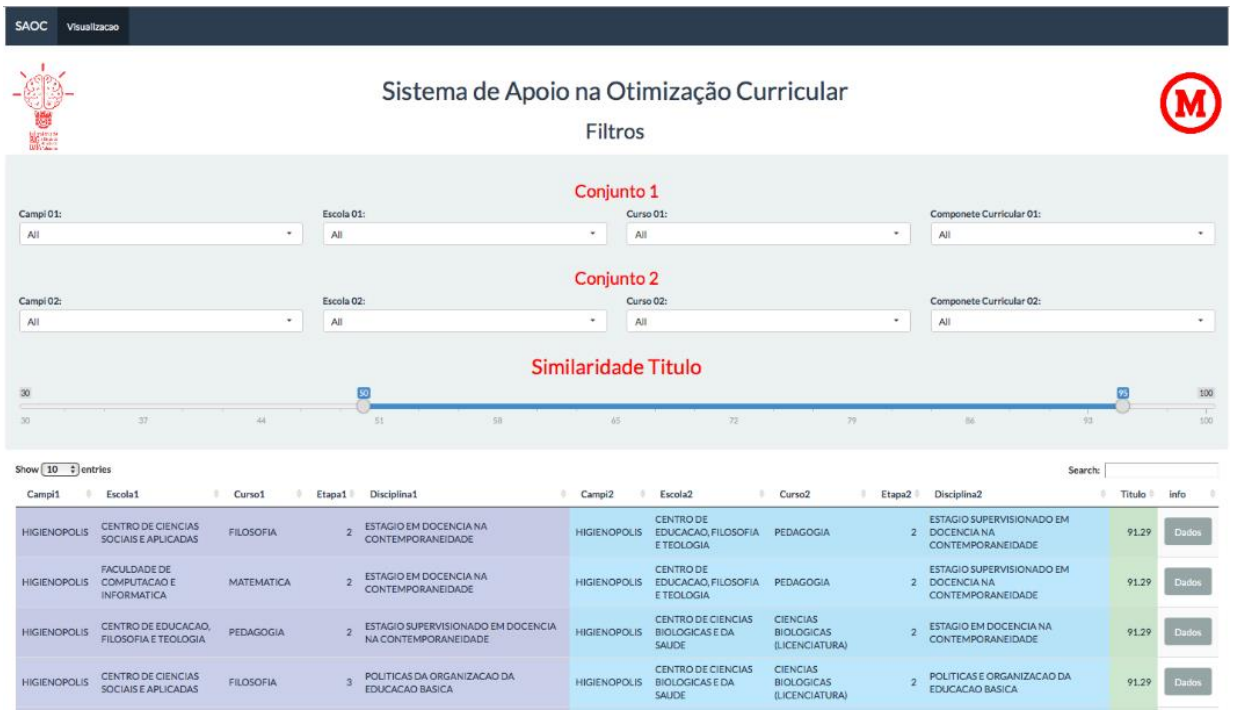

Figura 2 - Interface SaOC

Os dados disponíveis para análise estão estruturados como segue: eixo principal do componente curricular, a etapa a qual o eixo pertence, o nome do componente curricular, a quantidade de créditos demandadas distribuídas entre aulas presenciais, aulas em laboratório e aulas EAD bem como informações sobre o curso pertencente junto à escola em questão.

A padronização e unificação dos componentes curriculares é necessária para que, na elaboração de tais grades curriculares (junção dos componentes curriculares para a criação de um curso específico), possam existir componentes curriculares distintos. Os componentes curriculares analisados das 12 Faculdades estão contidos em diferentes tabelas, onde cada uma possui seu estilo próprio de agrupamento das informações, e dada tal situação, optou-se neste trabalho por padroniza-las de tal forma a conter apenas as informações relevantes à avaliação de similaridade entre esses componentes.

A padronização foi elaborada de modo a conter a escola origem do componente curricular, o curso no qual será ministrado, o nome do componente curricular e a quantidade de créditos distribuídas entre aulas presenciais, aulas em laboratório, aulas EAD pertinentes ao CC (Componentes Curriculares). A estruturação da solução foi dada através da representação de uma arquitetura em composta por dois módulos independentes.

A primeira etapa do processo contém a funcionalidade de envio de um arquivo parametrizado e seu processamento em linguagem R; enquanto a segunda etapa consiste na visualização dos dados filtrados e processados através da interface gráfica.

No processo de envio e processamento das informações necessárias tem sua implementação feita de forma independente à visualização, uma vez que tal processamento acontecerá de maneira sazonal e demandando um certo tempo, o qual não poderá ser tomado a cada vez que a visualização da interface gráfica for solicitada. Para tal, a primeira etapa do processo carregará a informação necessária, processará os dados e armazenará os resultados do processamento em um outro arquivo parametrizado (csv, por exemplo), o qual será salvo em uma estrutura de arquivos condizente com a projetada para execução do servidor R.

Para a realização do tratamento dos dados selecionados, onde chamamos de pré-processamento, utilizou-se a representação vetorial em que cada entrada é tida como um documento distinto, gerando assim um *corpus* onde os elementos possam ser organizados, colocando-os em uma estrutura de dados adequada para a aplicação de outras técnicas voltadas ao tratamento das informações.

Na sequência, a estrutura do *corpus* foi alterada para que os dados contidos sejam passados para letra maiúscula, armazenando o resultado em uma nova variável. Com base nessa primeira transformação, executou-se mais duas transformações, mas dessa vez removendo toda e qualquer pontuação existente e todos os números contidos no interior de cada documento do corpus, finalizando assim a etapa de análise lexical.

Com base nisso, deu-se início ao processo de eliminação de termos irrelevantes, chamados de *stopwords.* Os termos *stopwords* contidos no idioma português foram removidos dos documentos. E, por fim, iniciou-se o processo de redução dos termos ao radical.

A partir disso, a variável corpus já está preparada para ter sua representação vetorial gerada. A representação é baseada na frequência dos documentos com relação a

seus termos onde os documentos e termos estão alocados, permitindo que seja realizada uma seleção dos termos envolvidos baseando-se na quantidade mínima de caracteres e na frequência mínima, retornando assim uma matriz de frequência de termos e documentos.

Para ponderar a frequência de tais ocorrências dos termos nos documentos, utilizamos a frequência dos termos (*tf*) e o inverso da frequência nos documentos (*idf*), que utilizará como parâmetro a matriz de frequência de termos e documentos, resultando em uma matriz binária da ocorrência dos termos dados em um documento. A partir dessa etapa do processo, os dados carregados foram pré-processados e disponíveis para serem analisados através de algum algoritmo de mineração de dados como os de classificação, predição e/ou agrupamento, ou até mesmo a utilização de procedimentos de análises simplificadas para que a informação e o valor agregado aos dados textuais possam ser visualizados.

Para facilitar a manipulação dos resultados gerados, a matriz binária foi rotacionada para uma matriz de documentos por termos, ou seja, a dimensão será composta pelo número de documentos e número de termos. Por fim uma matriz foi gerada para o armazenamento do resultado referente à similaridade entre os documentos de maneira única, tendo então como dimensão o número de documentos.

A partir dessa matriz de similaridade é que o S*a*OC fará sua consulta e que posteriormente associara ao resultado as informações referentes à escola, curso, etapa e componente curricular.

Com todo o processamento acima realizado, utilizou-se o pacote "*Shiny*", que se trata de um framework de aplicação web para R, transformando as análises realizadas em aplicações web interativas, não requerendo nenhum conhecimento em HTML, CSS ou JavaScript, sendo assim, toda parte de visualização foi gerada através desse framework.

Para a análise dos resultados de similaridade utilizou-se o conceito de Precisão por Recuperação [MANNING, 2008]. Este é um conceito de Recuperação de Informação cuja utilização, como sugerida por CREIGHTON UNIVERSITY (2017), deve-se analisar como resposta de um processo de busca (neste trabalho a resposta das disciplinas similares) os documentos (disciplinas apresentadas aos pares) relevantes e irrelevantes. Sendo que um documento é relevante se satisfaz a informação procurada, ou seja, neste trabalho se as disciplinas são parecidas segundo a análise de um especialista. O documento será considerado como não relevante se as disciplinas apresentadas não forem semelhantes. Note que estas respostas são binárias e a contagem dos resultados permite que se tenha a precisão e a recuperação analisada.

A precisão (Pc) indica a quantidade de documentos recuperados e que são relevantes.

$$Pc = \frac{(itens\ relevantes\ recuperados)}{(itens\ recuperados)} \quad (3)$$

Equação 3 – Precisão

A recuperação (Rc) é a quantidade de documentos relevantes que são recuperados.

$$Rc = \frac{(itens\ relevantes\ recuperados)}{(itens\ relevantes)}$$

Equação 4 – Recuperação

Na seção seguinte apresenta-se os resultados experimentais, comparando as medidas Cosseno e Euclidiana por meio da *Pc* e *Rc*.

5. Resultados

Com a implementação do S*a*OC, após o processamento exaustivo guiado pelas técnicas supracitadas, obtemos uma matriz de igualdade, composta por dois blocos de informações, onde cada um contém as informações (“Campi”, “Escola”, “Curso”, “Etapa” e “Disciplina”) referentes à disciplina que será comparada com as informações do outro bloco, contendo também à similaridade entre os títulos das disciplinas em questão que foram comparadas no dado momento.

Para avaliar a precisão e recuperação do sistema, foi inserido no processo o usuário do S*a*OC e que é especialista na análise de CC da Universidade, o qual faz a otimização dessas CC de forma manual. O critério de consulta para a análise baseou-se na frequência dos termos mais relevantes entre todos os documentos, destacando então as palavras “Direito”, “Elemento” e “Matemática”. Portanto, com base a cada 10 documento recuperados, avaliou-se quais são efetivamente relevantes e não relevantes, comparando então a efetividade da distância cosseno e a distância euclidiana como medida de similaridade. Os resultados estão apresentados na Figura 3.

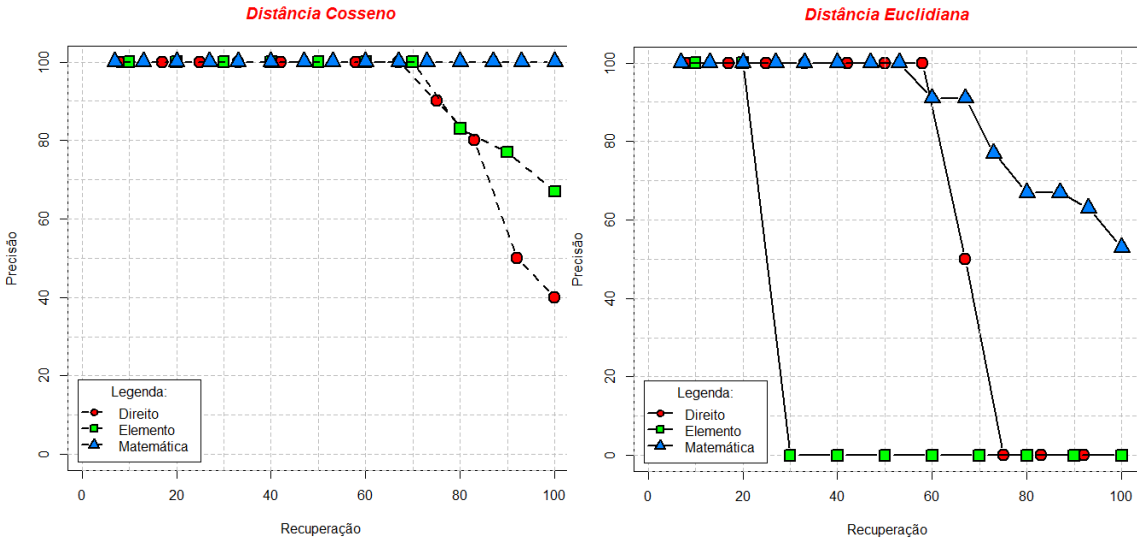

Figura 3 - Precisão por recuperação com base na distância cosseno e Euclidiana.

Os resultados obtidos e representados na Figura 3 indicam que a distância cosseno tem melhor manutenção da precisão, mesmo quando há aumento da recuperação em relação à distância euclidiana. Portanto, a distância cosseno deve ser adotada como a métrica para o sistema apresentado na Figura 1.

6. Conclusão

Para resolver o problema principal voltado a difícil aplicação do conceito pluridisciplinar em uma universidade, foram utilizadas técnicas de mineração, propondo um processo que coleta as Componentes Curriculares de Faculdades de uma Universidade (seleção), faz o tratamento dos nomes das CC (pré-processamento, tokenização, *stopwords* e *stemming*), permitindo a representação numérica desses nomes (representação). Com as CC representadas, aplica-se medidas de similaridade como distância Cosseno e Euclidiana (similaridade), a fim de fazer descobertas sobre as disciplinas com nomes parecidos; e, por fim, apresenta-se as CC similares, permitindo ainda que se visualize outras informações das mesmas como carga horária, formato de disciplinas e Faculdade que a oferece (visualização).

A evidencia da melhor manutenção de precisão se deu através da aplicação da distância cosseno. A implementação de um sistema que otimiza o uso comum de CC, tomando como base principal à similaridade entre o nome das disciplinas de uma Universidade.

Neste trabalho utilizou-se apenas os nomes das Componentes Curriculares e pretende-se, como trabalhos futuros, ampliar a comparação para as ementas e referências bibliográficas das disciplinas, trazendo assim uma maior assertividade de comparação para o modelo proposto.

Referências

BADAWY, Mohammed; EL-AZIZ, A. A. Abd; HEFNY, Hesham A.. Analysis of learning objectives for higher education textbooks using text mining. Computer Engineering Conference (ICENCO), Cairo, Egypt, n. 12, p. 202-207, dez. 2016.

CREIGHTON UNIVERSITY. Measuring search effectiveness. Disponível em: <https://www.creighton.edu/fileadmin/user/hsl/docs/ref/searching_-_recall_precision.pdf>. Acesso em: 12 mai. 2017.

GOMAA, Wael H.; FAHMY, Aly A.. A Survey of Text Similarity Approaches. International Journal of Computer Applications, Cairo, Egypt, v. 68, n. 13, p. 13-18, abr. 2013.

IDA, Masaaki. Textual Information and Correspondence Analysis in Curriculum Analysis. FUZZ-IEEE, Korea, v. 09, p. 666-669, ago. 2009.

KAWINTIRANON, K. et. al. Understanding Knowledge Areas in Curriculum through Text Mining from Course Materials. IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Bangkok, Thailand, p. 161-168, dez. 2016.

LAHITANI, Alfirna Rizqi; PERMANASARI, Adhistya Erna; SETIAWAN, Noor Akhmad. Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment. 4th International Conference on Cyber and IT Service Management, Indonesia, n. 04, abr. 2016.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHUTZE, Hinrich. Introduction to information retrieval. 1 ed. New York, USA: Cambridge University Press, 2008. 124-161 p.

SILVA, Leandro Augusto Da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. Introdução à mineração e dados: Com aplicações em R. 1 ed. Rio de Janeiro, Brasil: Elsevier, 2016. 29-73 p.

ZHANG, T. et. al. Document Clustering in Correlation Similarity Measure Space. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, v. 24, n. 6, p. 1002-1013, dez. 2016.

ZHOU, Bing; YAO, Yiyu. Evaluating Information Retrieval System Performance Based on User Preference. Journal of Intelligent Information Systems, Regina, Saskatchewan, Canada, v. 34, n. 3, p. 227–248, jun. 2010.

PEZZINI, Anderson. MINERAÇÃO DE TEXTOS: CONCEITO, PROCESSO E APLICAÇÕES. Revista Eletrônica do Alto Vale do Itajaí, Santa Catarina, v. 5, n. 8, p. 1-13, dez. 2016.

OLIVEIRA, João Ferreira De; CATANI, Afranio Mendes. A educação superior. ResearchGate, [S.L], abr. 2015. Disponível em: <https://www.researchgate.net/publication/268430084_A_Educacao_Superior>. Acesso em: 17 jul. 2017.,