ISSN 9876-8675

Conversor XML para TXT: ferramenta para análise de presença social associada a clustering no *software Eurekha*!

Yasmmin C. Martins, Breno F. T. Azevedo, Hélvia P. P. Bastos

Núcleo de Informática na Educação (NIE)
Instituto Federal Fluminense Campus Campos-Centro (IFF) – Campos dos Goytacazes, RJ – Brasil
{helviabastos, brenoter}@gmail.com, nim_asay@hotmail.com

Abstract: The paper presents a tool able to reconfigure XML files resulting from the automatic analysis of text indicators of social presence performed by the software Presente! (Kambara-Silva, 2011). For clustering experiments in the software Eurekha! (Wives, 2004), these XML files must be converted into TXT format. The converter described in the study enables such files to be graphically analyzed, thus helping in decision-making processes.

Resumo: O trabalho apresenta uma ferramenta que realiza a reconfiguração de arquivos XML resultantes do processo de análise automática de marcas textuais de presença social realizado pelo software Presente! (Kambara-Silva, 2011). Para experimentos de agrupamento no software Eurekha! (Wives, 2004), os arquivos XML precisam ser formatados em TXT. O conversor descrito neste estudo possibilita que esses arquivos possam ser analisados graficamente, fato que auxilia a tomada de decisão por parte do analista.

1. Introdução

Este trabalho resulta de atividades de pesquisa realizadas para o projeto "Mineração de Dados Textuais Eletrônicos" (Núcleo de Informática na Educação – NIE). A mineração de textos (MT) pode ser definida como a extração de informações úteis de uma série de documentos textuais. Constituindo uma vasta área de estudos, a MT tem aplicação em diversos campos, inclusive a busca de dados textuais em bases eletrônicas de apoio a projetos pedagógicos.

ISSN 9876-8675

O objetivo do artigo é apresentar uma ferramenta desenvolvida para otimizar tarefas de clustering (agrupamento) realizada pelo programa Eurekha! [Wives 2004]. O agrupamento de textos (clustering) é uma técnica utilizada para agrupar documentos semelhantes. No agrupamento, os documentos podem aparecer em vários subtópicos nos resultados da pesquisa. Um algoritmo de agrupamento básico cria um vetor de tópicos para cada texto, e calcula os pesos para identificar em qual grupo um documento deve fazer parte [Gupta e Lehal, 2009; Kobayashi e Aono, 2004].

O experimento de conversão descrito neste estudo utilizou arquivos no formato XML resultantes do processo de mineração feito pelo software Presente! [Kambara-Silva 2011]. Esse programa foi desenvolvido para identificar pistas textuais de presença social em ambientes virtuais de aprendizagem conforme modelo de Bastos [2012]. O objetivo desse processo computacional é auxiliar o docente no acompanhamento da participação dos alunos em cursos a distância. As diferentes partes que compõem o programa, seus objetivos e funcionamento são descritas na Seção 2.

A Seção 3 descreve como a ferramenta criada na pesquisa realiza a conversão de arquivos XML para TXT, possibilitando uma verificação e visualização mais correta e objetiva dos dados de agrupamento feitos no Eurekha!.

2. Software Presente!

A principal função do software Presente! é fazer análise automática de marcas linguísticas de presença social (PS), segundo as classes designadas no Modelo Presença Plus [Bastos 2012]. Presença social pode ser compreendida como o grau de sentimento, percepção e reação dos indivíduos ao interagirem por meio de recursos de comunicação mediada por computador [Tu e McIsaac 2002].

Para identificar essas pistas textuais nas postagens em fóruns e chats, [Kambara-Silva 2011] desenvolveu softwares auxiliares para gerar alguns arquivos de entrada: o conversor de HTML para XML, o construtor de categorias/gerador de classes.

ISSN 9876-8675

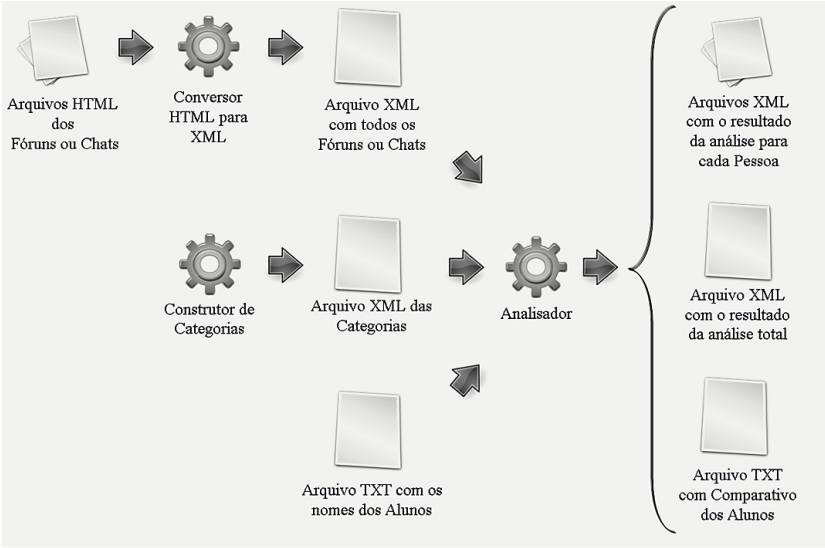


Figura 1: Esquema de funcionamento do Analisador de Presença Social [Kambara-Silva 2011]

A figura acima ilustra a ordem e que tipos de arquivos são analisados pelas ferramentas auxiliares para originar arquivos de resultado. Essas ferramentas são explicadas a seguir.

2.1. Conversor HTML para XML

Este software transforma as páginas web de fóruns e chats, encontradas em determinado ambiente virtual de aprendizagem, em documentos estruturados que possuem somente as tags necessárias para uma análise posterior. Essa operação é feita pela retirada somente de informações como o nome do autor da postagem, o assunto, a data, e o conteúdo da postagem.

O uso da página web completa seria inviável porque ela possui muitas informações que não possuem relação com o que pretende ser analisado, como tags de marcação de estética (como

ISSN 9876-8675

div e span), tags de imagem, as que indicam ênfase (negrito, cores e tamanhos de fonte diferentes), entre outras. O conversor prepara os arquivos essenciais para a análise correta. Essa etapa é seguida da geração de arquivo XML com todos os posts e informações necessárias contidas neles.

2.2. Gerador de classes

Esse software é utilizado para definir as pistas ou palavras/expressões que podem estar no texto dos posts, classificando-as em relação a uma determinada classe e subclasse de PS. O Modelo PPlus possui quatro classes de indicadores de PS: afetividade, interatividade, coesão social e força. Cada classe apresenta subclasses, entre elas: emoção, humor, termos de sociabilidade e ênfase.

Pode-se retirar essas possibilidades de um dicionário da língua que será utilizada na análise, mas esse uso é opcional. Quanto mais completa e detalhada for a classificação das palavras e expressões nas categorias acima, mais precisa será a análise e, provavelmente, os resultados e as interpretações acerca das intenções e participação do aluno no ambiente virtual.

Nessa etapa também é gerado um arquivo XML com o mapeamento de todas as categorias e subcategorias, com suas respectivas pistas textuais.

Após obter os arquivos necessários nos softwares citados, alimenta-se o software denominado Analisador, gerando, assim, um arquivo XML relativo a cada estudante, com a quantidade de pistas encontradas para cada categoria e subcategoria com base nos posts realizados por esse aluno. A Figura 2 apresenta como esses arquivos são visualizados.

ISSN 9876-8675

```
<?xml version="1.0" encoding="UTF-8" ?>

<classes>
    <quantidade>18</quantidade>
    <afetividade>
        <quantidade>5</quantidade>
        <emocao>
            <quantidade>2</quantidade>
            <interjeicoes_e_expressoes_interjetivas>
                <quantidade>1</quantidade>
                <expr_simples>
                    <quantidade>1</quantidade>
                </expr_simples>
                <inicio_da_frase>
                    <quantidade>0</quantidade>
                </inicio_da_frase>
            </interjeicoes_e_expressoes_interjetivas>
            <onomatopeias>
                <quantidade>0</quantidade>
            </onomatopeias>
            <emoticons_e_gifs>
                <quantidade>0</quantidade>
            </emoticons_e_gifs>
            <pontuacao_repetida>
                <quantidade>0</quantidade>
            </pontuacao_repetida>
```

Figura 2: Ilustração de um dos arquivos XML gerados por aluno

3. Metodologia

Uma das formas de se organizar e analisar o conteúdo dos arquivos XML é fazer sua classificação em grupos que possuem algo em comum. Isso é feito a partir de algoritmos que fazem correspondências métricas e análises qualitativas em relação aos documentos textuais de entrada, como ocorre no software de mineração e agrupamento Eurekha!.

Como este software de agrupamento não possui compatibilidade com arquivos XML, a princípio, foi feita apenas a troca da extensão para texto, mantendo as tags com os nomes e os números das quantidades. Entretanto, o software não fornece relevância a números, que no caso seria a principal informação que influenciaria no agrupamento.

Os algoritmos de classificação utilizam principalmente palavras. Para fazer o software interpretar corretamente a informação que se quer usar como parâmetro de classificação – as quantidades numéricas das classes e subclasses, o processo de transformação foi baseado na repetição dos nomes das mesmas de acordo com suas quantidades.

ISSN 9876-8675

Assim, o algoritmo continuará fazendo a classificação atribuindo peso às palavras. No entanto, como as palavras agora correspondem às quantidades, ao comparar os arquivos, a classificação ocorrerá de forma adequada.

```
<?xml version="1.0" encoding="UTF-8" ?>
<classes>
    <quantidade>18</quantidade>
    <afetividade>
        <quantidade>5</quantidade>
        <emocao>
            <quantidade>2</quantidade>
            <interjeicoes_e_expressoes_interjetivas>
                <quantidade>1</quantidade>
                <expr_simples>
                    <quantidade>1</quantidade>
                </expr_simples>
```

Figura 3: Trecho do arquivo XML antes de passar pela ferramenta

```
afetividade afetividade afetividade afetividade afetividade
emoçao emoção
interjeicoes_e_expressões_interjetivas
expr_simples
```

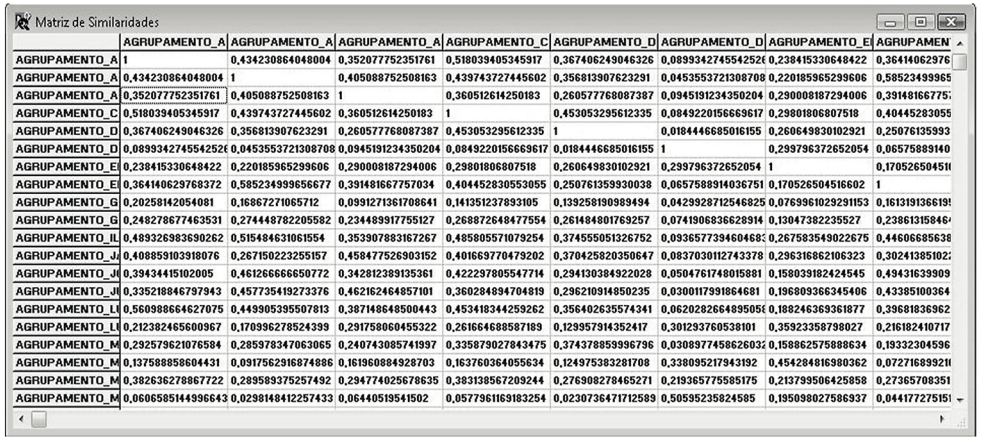
Figura 4: O mesmo trecho acima após ser transformado pela ferramenta

As Figuras 3 e 4 apresentam um trecho do documento XML após passar pela ferramenta de pre paração. Assim, eles podem ser interpretados e agrupados corretamente pelo Eureka!

ISSN 9876-8675

4. Resultados

Após obter os arquivos no analisador de presença social, foram carregados os arquivos XML resultantes já preparados pela ferramenta. Para fazer o agrupamento, basta escolher a opção "identificar relacionamentos". Para realizar essa operação foi escolhido o algoritmo Best-star (que foi o que melhor classificou os grupos). Desse modo, logo se visualiza uma matriz de similaridades, conforme a figura 5.



	AGRUPAMENTO_A	AGRUPAMENTO_A	AGRUPAMENTO_A	AGRUPAMENTO_C	AGRUPAMENTO_D	AGRUPAMENTO_D	AGRUPAMENTO_E	AGRUPAMEN
AGRUPAMENTO_A	1	0,434230864048004	0,352077752351761	0,518039405345917	0,367406249046326	0,0899342745542526	0,238415330648422	0,36414062976
AGRUPAMENTO_A	0,434230864048004	1	0,405088752508163	0,439743727445602	0,356813907623291	0,0453553721308708	0,220185965299606	0,585234999656
AGRUPAMENTO_A	0,352077752351761	0,405088752508163	1	0,360512614250183	0,260577768087387	0,0945191234350204	0,290008187294006	0,391481667757
AGRUPAMENTO_C	0,518039405345917	0,439743727445602	0,360512614250183	1	0,453053295612335	0,0849220156669617	0,298018068607518	0,404452830553
AGRUPAMENTO_D	0,367406249046326	0,356813907623291	0,260577768087387	0,453053295612335	1	0,0184446685016155	0,260649830102921	0,250761359933
AGRUPAMENTO_D	0,0899342745542526	0,0453553721308708	0,0945191234350204	0,0849220156669617	0,0184446685016155	1	0,299796372652054	0,065758891403
AGRUPAMENTO_E	0,238415330648422	0,220185965299606	0,290008187294006	0,298018068607518	0,260649830102921	0,299796372652054	1	0,170526504516
AGRUPAMENTO_E	0,36414062976372	0,585234999656677	0,391481667757034	0,404452830553055	0,250761359930038	0,0657588914036751	0,170526504516602	1
AGRUPAMENTO_G	0,20258142054081	0,16867271065712	0,0991271361708641	0,141351237893105	0,139258190989494	0,0429287125468251	0,0769961029291153	0,161319136619
AGRUPAMENTO_G	0,248278677463531	0,274448782205582	0,234489917755127	0,268872648477554	0,261484801763257	0,0741906836628914	0,13047382235527	0,238613158464
AGRUPAMENTO_IL	0,489326983690262	0,515484631061554	0,353907883167267	0,485805571079254	0,374555051326752	0,0936577394604683	0,267583549022675	0,446066856381
AGRUPAMENTO_J	0,408859103918076	0,267150223255157	0,458477526903152	0,401669770479202	0,370425820350647	0,0837030112743378	0,296316862106323	0,302413851022
AGRUPAMENTO_J	0,39434415102005	0,461266666650772	0,342812389135361	0,422297805547714	0,294130384922028	0,0504761748015881	0,158039182424545	0,434316399095
AGRUPAMENTO_J	0,335218846797943	0,457735419273376	0,462162464857101	0,360284894704819	0,296210914850235	0,0300117991864681	0,196809366345406	0,433851003644
AGRUPAMENTO_L	0,560988664627075	0,449905395507813	0,387148648500443	0,453418344259262	0,356402635574341	0,0620282664895058	0,188246369361877	0,396818369624
AGRUPAMENTO_L	0,212382465600967	0,170996278524399	0,291758060455322	0,261664688587189	0,129957914352417	0,0301293760538101	0,359233587980277	0,216182410717
AGRUPAMENTO_M	0,292579621076584	0,285978347063065	0,240743085741997	0,335879027843475	0,374378859996796	0,0308977458626033	0,158862575888634	0,193323045962
AGRUPAMENTO_M	0,137588858604431	0,091756291687488 6	0,161960884928703	0,163760364055634	0,124975383281708	0,0380952179431 92	0,454284816980362	0,072716899211
AGRUPAMENTO_M	0,382636278867722	0,289589375257492	0,294774025678635	0,383138567209244	0,276908278465271	0,0219365775585175	0,213795906425858	0,273657083513
AGRUPAMENTO_M	0,0606585144996643	0,0298148412257433	0,064405195415 02	0,0577961169183254	0,0230736471712589	0,0059523582458 5	0,195098027586937	0,044177275151

Figura 5: Representação da matriz de similaridades gerada no experimento

Essa matriz é gerada através da comparação entre os conteúdos de todos os arquivos carregados. Por isso, quando são comparados os mesmos arquivos, o grau de igualdade é máximo. Para os diferentes, podem-se ter resultados entre 0.03% e 52% aproximadamente, ou seja, podem-se ter as mesmas palavras, nas mesmas quantidades ou não. Quanto mais diferentes forem, mais próximos de zero estarão.

Esta operação pode ser utilizada para saber quais alunos possuem um grau de envolvimento e participação semelhante. No entanto, esta informação seria generalizada, pois até então não

ISSN 9876-8675

foram identificadas as diferenças nos conteúdos e sim as informações quantitativas gerais dos arquivos.

Outro resultado que o software provê é o gráfico de colunas com o percentual relativo de documentos em cada cluster (grupo). Através deste gráfico, pode ser percebida a principal diferença

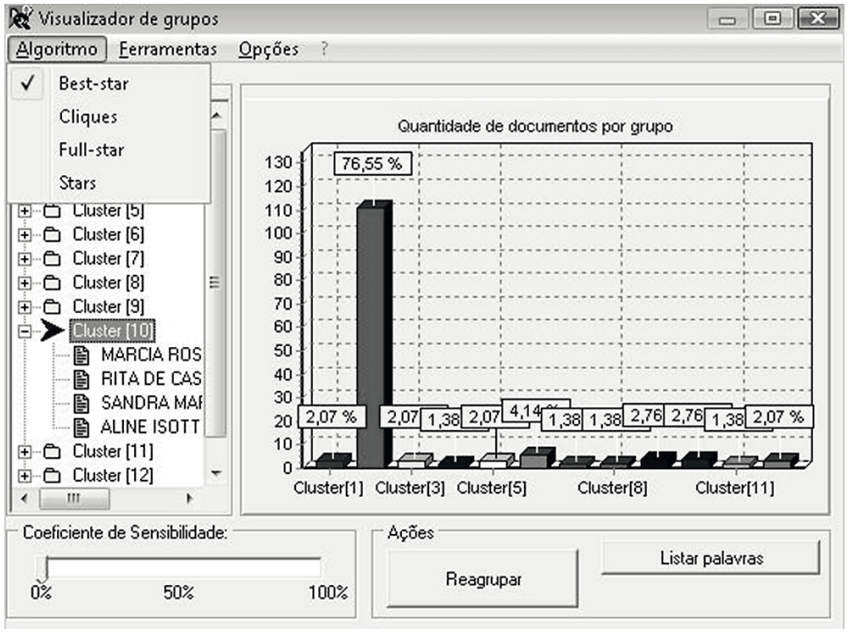


Figura 6: Experimento de clustering sem passar os arquivos na ferramenta

Na figura 6 se percebe que houve um agrupamento não muito compreensível e coerente com os números e as divergências nos conteúdos dos arquivos. No grupo em verde, que corresponde a aproximadamente 77% dos arquivos basicamente, só havia os arquivos que continham zero ocorrência de quantidade, ou seja, praticamente não houve participação desses alunos. Nos

ISSN 9876-8675

outros casos, não houve um padrão coerente de classificação ao visualizar as características dos arquivos que estavam num mesmo grupo.

Na figura 7, verifica-se que houve diferenças menos contrastantes e expressivas em relação aos percentuais de arquivos nos grupos. De acordo com a nova configuração e distribuição da informação nos arquivos, os grupos apresentaram uma lógica de explicação na sua classificação.

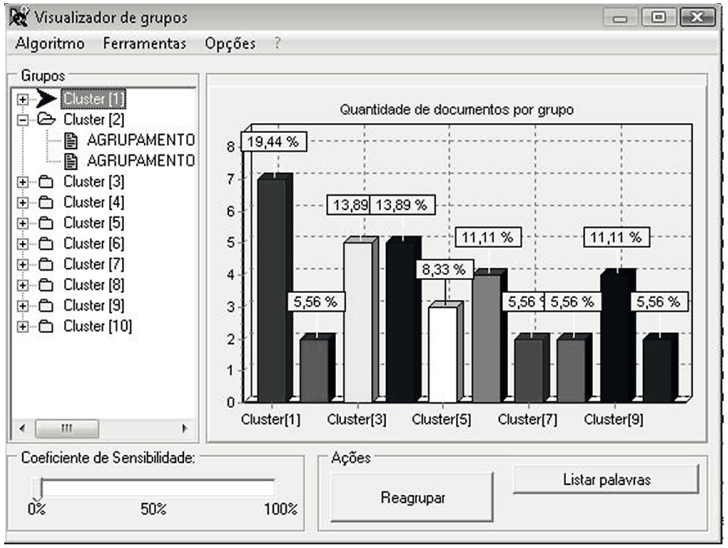


Figura 7: Experimento utilizando a ferramenta nos arquivos de entrada

Para saber o que há de comum nos arquivos que pertenceram a um mesmo grupo, assim como o que há de diferente entre eles e os dos demais grupos, basta executar a ação "Listar palavras". O software mostrará que naquele grupo as palavras mais relevantes possuem suas frequências em igual quantidade.

ISSN 9876-8675

Dependendo das palavras com mais ocorrências e mais relevância no conjunto, pode-se descobrir a classificação do arquivo do aluno (em qual grupo, classe e subclasse ele mais se caracterizou). Para analisar as diferenças internas dos grupos foram utilizados 3 dos 10 grupos que apresentaram percentuais distintos (os grupos 1, 2 e 4).

4.1. Interpretação da classificação do grupo 1

A Figura 8 apresenta a lista das palavras com ocorrências e relevâncias iguais para os arquivos que a este grupo pertencem. Neste grupo, pode-se verificar que houve ocorrências consideráveis de classes e subclasses diferentes.

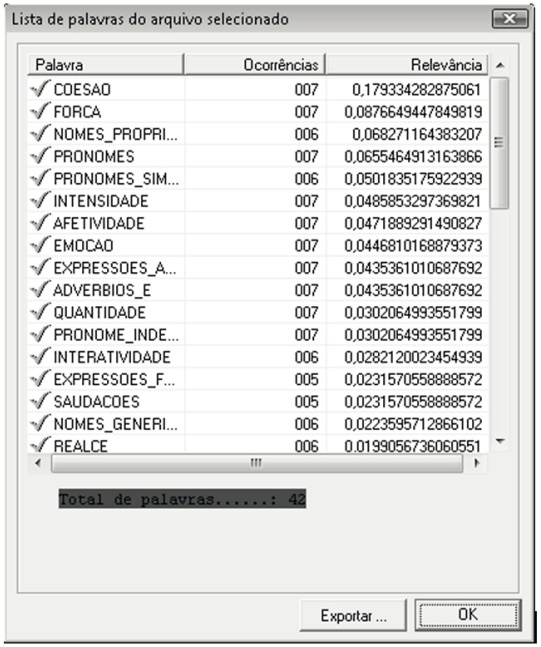


Figura 8: Perfil de classificação do grupo/cluster 1

ISSN 9876-8675

A ordem das palavras está de acordo com a ordem decrescente do seu nível de relevância em relação a todas as palavras encontradas. Por isso, como a subclasse de nomes próprios foi considerada mais relevante nos textos, mesmo tendo uma frequência maior que intensidade, por exemplo, ele ficou acima deste. Destaca-se que neste grupo foram encontradas 42 palavras.

4.2. Interpretação da classificação do grupo 2

A figura 9 mostra a lista de palavras do grupo 2, comuns a todos os arquivos que a ele pertencem. Pode-se perceber que apesar da mistura de classes e subclasses, as ocorrências foram parecidas. No entanto, elas foram muito menores que as do grupo anterior.

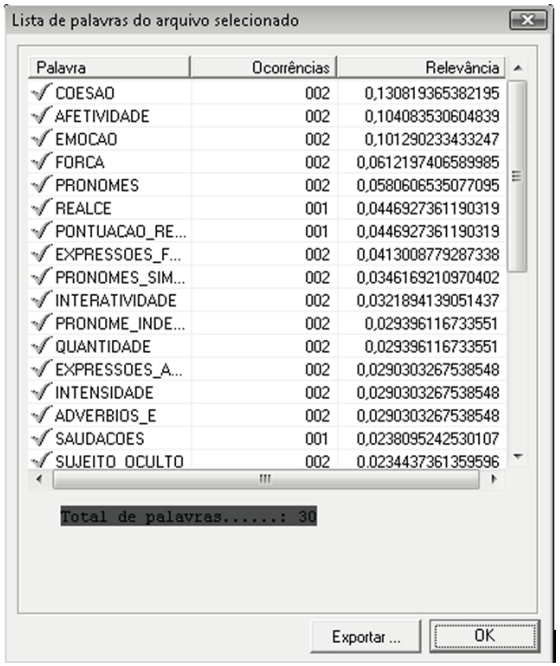


Figura 9: Perfil de classificação para o grupo/cluster 2.

ISSN 9876-8675

Este grupo tem em comum o fato de que as classes de PS mais encontradas para os alunos foi o das categorias coesão, afetividade, emoção etc., com no máximo 2 ocorrências. Obtiveram-se também ocorrências de outros tipos de PS. Elas foram parecidas entre si, porém com valores pequenos. Nesse caso, os arquivos continham 30 palavras em comum nesse grupo.

4.3. Interpretação da classificação do grupo 4

A figura 10 mostra a lista de palavras do grupo 4, também com o mesmo comportamento para todos os arquivos do grupo. Pode-se perceber que houve ocorrências de poucos tipos de classes e subclasses. Algumas destas apresentaram um número considerável, enquanto outras obtiveram uma quantidade bem baixa.

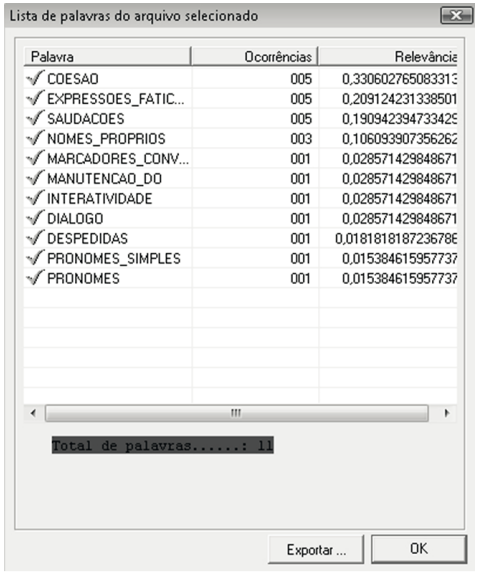


Figura 10: Perfil de classificação do grupo/cluster 4.

ISSN 9876-8675

Nesse grupo houve poucas classes/subclasses diferentes encontradas. As que foram encontradas obtiveram uma frequência considerável, principalmente nas de coesão, expressões fáticas e saudações. As outras classes/subclasses encontradas apresentaram um número bem menor. Nesse caso, existiam somente 11 palavras em comum.

Outro fator importante é que, antes da ferramenta, os documentos que continham zero ocorrência também foram processados. No entanto, após passar pelo programa eles desapareceram, pois não podiam ser repetidos.

A ferramenta pode ser usada como apoio na análise, interpretação e apresentação de resultados mais claros para os utilizadores do Analisador de PS, auxiliando na tomada de decisões a partir da visualização do comportamento dos grupos.

5. Conclusão

Após os experimentos percebe-se que a ferramenta auxiliou de forma significativa na compreensão dos perfis de comportamento dos alunos, ao se comunicarem através de ambientes virtuais de aprendizagem. Isso demonstra que o software é uma possível solução para melhorar a apresentação dos resultados e a tomada de decisões por parte do usuário, quando lidar com arquivos no formato XML no software Eurekha!.

Referências

Bastos, H. P. P. (2012). Presença Plus: modelo de identificação de presença social em ambientes virtuais de ensino e aprendizagem. Tese de doutorado. Programa de Pós-Graduação em Informática na Educação. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS.

Gupta, V.; Lehal, G. S. (2009) "A Survey of Text Mining Techniques and Applications". Journal of Emerging Technologies in Web Intelligence, v. 1, n. 1.

ISSN 9876-8675

Kambara-Silva, J.K. (2011) "Automatização do processo de aquisição de presença social em fóruns e chats". Trabalho de Conclusão de Curso. Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS

Kobayashi, M.; Aono, M. (2004) "Vector Space Models for Search and Cluster Mining". In: BERRY, M. W. (ed.). Survey of text mining: clustering, classification, and retrieval. New York: Springer-Verlag. p. 103-122.

Tu C.H; McIsaac (2001) "The relationship of social presence and interaction in online classes". In The American Journal of Distance Education, v.16, n.3, p. 131-50.

Wives, L. K. (2004). Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos. Exame de Qualificação (doutorado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.