

Agrupamento de Informações Textuais

Leandro Krug Wives
José Palazzo M. De Oliveira
wives@inf.ufrgs.br

Resumo

Este artigo descreve a aplicação de técnicas de agrupamento de objetos em informações textuais, visando amenizar alguns dos problemas causados pela enorme quantidade de informações disponível na atualidade. Utilizando-se de uma ferramenta de agrupamento, o usuário consegue manipular mais facilmente as informações que lhe interessam, descartando aqueles grupos que não são de seu interesse. No decorrer do texto, encontram-se descritas a metodologia de agrupamento adotada, a ferramenta de agrupamento implementada e algumas conclusões sobre os resultados parciais obtidos nos testes de agrupamento de documentos realizados até o momento.

1. INTRODUÇÃO

Com o advento da Internet o volume de informações que uma pessoa tem acesso cresce diariamente. Porém, este volume crescente de informações torna cada vez mais difícil a tarefa de assimilação da informação por um ser humano. Em estudo apresentado em trabalho anterior [1], verifica-se que uma quantidade muito grande de informações, desordenadas, pode levar uma pessoa a um problema denominado *sobrecarga de informações*.

A *sobrecarga de informações*, segundo [2], ocorre quando uma pessoa, ao realizar uma busca por informações, obtém um número excessivo de informações (mesmo que relevantes) e não consegue absorvê-las ou tratá-las. Este trabalho possui como foco principal este problema, e pode ser enquadrado na área de *recuperação de informações* – área cuja comunidade dedica grande interesse e esforços na solução de problemas relacionados a manipulação de informações.

A *sobrecarga de informações* pode ocorrer em diversos locais. Podem ser citados o serviço de correio eletrônico e a localização de páginas WEB. No primeiro, correio eletrônico, é comum que uma pessoa (o usuário) receba uma grande quantidade de informações diariamente. Estas informações, contidas em mensagens, podem ser provenientes dos mais variados locais, com os mais variados assuntos. Do mesmo modo, as páginas WEB também estão dispostas de maneira desorganizada, necessitando que o usuário analise-as exaustivamente até que encontre a informação de que necessita (mesmo que utilize uma ferramenta de recuperação de informações como o Altavista®). O usuário, conseqüentemente, perde muito tempo tentando organizar estas informações para que possa assimilá-las corretamente ou para que possa encontrá-las mais rapidamente no futuro.

Deste modo, este trabalho sugere a utilização de um sistema capaz de agrupar automaticamente as informações para o usuário, eliminando ou minimizando o problema da *sobrecarga de informações*. Nas páginas seguintes apresenta-se uma metodologia básica de agrupamento de objetos textuais, com base nas técnicas de agrupamento encontradas na literatura da área. Após, encontra-se uma breve descrição sobre a implementação de algumas técnicas de agrupamento, utilizando a metodologia apresentada, assim como as conclusões sobre testes realizados até o momento.

2. AGRUPAMENTO DE INFORMAÇÕES TEXTUAIS

Para que os documentos textuais possam ser agrupados, é necessário utilizar-se de alguma técnica de agrupamento de objetos. Segundo [3], o *agrupamento de objetos* (clustering) é uma das técnicas utilizadas na *Descoberta de Conhecimento e Mineração de Dados* – áreas da *Inteligência Artificial*.

O processo de agrupamento automático de objetos baseia-se na hipótese de que objetos semelhantes tendem a permanecer em um mesmo *grupo* (cluster), pois possuem atributos em comum. Estes objetos, pertencentes ao mesmo grupo, tendem a ser relevantes ao mesmo assunto. Esta hipótese, segundo [4], é conhecida por *hipótese de agrupamento* (cluster hypothesis).

A técnica de agrupamento consiste em organizar uma série desorganizada de objetos em grupos de objetos similares. Este tipo de técnica é recomendada quando não há uma discriminação prévia de classes, sendo útil em casos onde não há a possibilidade de alocar um especialista na tarefa de separação de objetos em classes (manualmente). Por outro lado, mesmo quando há um especialista, a técnica também pode ser utilizada a fim de facilitar o trabalho do mesmo.

Porém, segundo [5], a maioria das técnicas de agrupamento atuam sobre dados estruturados, ou seja, aqueles dados convencionais, armazenados em Sistemas de Gerência de Bancos de Dados, mais fáceis de serem tratados por meios computacionais. Como os dados textuais não possuem estruturas (campos predefinidos), é necessário adaptar estas técnicas ao formato textual.

Devido a isso, novas técnicas de agrupamento foram (e ainda estão sendo) construídas para estes tipos de informação sem estrutura (informação textual). Porém, ainda não há um padrão estabelecido. A seção seguinte descreve a metodologia básica de agrupamento de objetos textuais descrita na literatura da área.

3. METODOLOGIA DE AGRUPAMENTO

Existem duas formas de se agrupar objetos. A primeira busca identificar hierarquias entre os grupos de elementos (objetos), sendo, portanto, aglomerativa. Neste caso, inicialmente, cada objeto é colocado em um grupo distinto. A seguir, os grupos são reagrupados em pares, de maior similaridade. Este processo é repetido até que o número de grupos torne-se suficientemente pequeno. Ao final, uma estrutura similar a uma árvore é obtida.

A segunda forma de agrupamento é obtida particionando-se linearmente os objetos, constituindo grupos disjuntos, distintos e não hierárquicos (ou seja, não há ligações entre os grupos como no caso anterior). Os algoritmos utilizados neste trabalho enquadram-se neste último grupo. Esta forma de agrupamento foi adotada porque a utilização de grupos distintos facilita a aplicação de uma ferramenta de análise/descoberta de conhecimento, já que não há ligações entre os grupos. Além disso, o resultado de um algoritmo linear pode ser transformado em uma estrutura hierárquica, aplicando-se o método de agrupamento recursivamente. Isso significa que a construção de uma hierarquia pode ser considerada uma etapa adicional, que pode ser mais exigente em matéria de recursos computacionais.

De qualquer forma, independentemente da técnica adotada, segundo análise realizada na literatura da área, é possível concluir que as técnicas existentes agrupam os objetos realizando as seguintes etapas: etapa de *identificação e seleção de características* nos objetos, etapa de *cálculo de similaridades* entre os objetos e etapa de *agrupamento*, de acordo com os graus de similaridades encontrados.

3.1. IDENTIFICAÇÃO E SELEÇÃO DE CARACTERÍSTICAS

Como as informações (objetos) tratados possuem uma forma textual, as palavras contidas no textos são utilizadas como características destes objetos. Cada palavra é selecionada como uma característica, e quanto maior for sua frequência (quanto mais aparecer no documento) maior é seu grau de importância (neste documento). Teoricamente, existem outras formas de identificação e seleção de características. Algumas, por exemplo, desconsideram palavras pouco importantes (que apareçam muito pouco em um documento) e palavras que apareçam em muitos documentos (porque não discriminam os objetos). Muitas delas foram definidas com o objetivo de aumentar a velocidade de processamento dos algoritmos de agrupamento. Porém, neste artigo, não há necessidade de detalhá-las (maiores detalhes sobre algumas destas técnicas podem ser obtidos em [6]). O importante é salientar o fato de que há uma fase de identificação características, e que estas características possuem um grau de importância para o objeto.

3.2. IDENTIFICAÇÃO DOS GRAUS DE SIMILARIDADE

Tendo-se a informação das características dos objetos e seus respectivos graus de importância consegue-se identificar as características comuns a alguns documentos. Existem diversas *funções de similaridade*, algumas citadas em [6], porém, não há um consenso nem uma análise detalhada sobre qual delas é a mais eficiente. Deste modo, como na maioria dos casos, são considerados similares aqueles objetos que possuem o maior número de características em comum. Logo, a técnica adotada neste trabalho compara exaustivamente, através de operadores *fuzzy*, todas as características dos objetos, levando em conta, é claro, o seu grau de importância no objeto. Se a característica não for muito importante para um dos objetos ela influi negativamente no grau de similaridade, e vice-versa. Finalmente, tendo-se os graus de similaridade entre todos os objetos, constrói-se a matriz de similaridade entre os objetos, que é utilizada na etapa seguinte.

3.3. AGRUPAMENTO

Após terem sido concluídos os cálculos de similaridade, e, portanto, após ter sido criada a matriz de similaridades, é necessário estabelecer algum tipo de restrição (regra) que irá definir os grupos de objetos (*clusters*). É neste ponto que a grande maioria dos algoritmos de agrupamento se diferencia: na restrição. Como exemplo de restrição, um dos algoritmos mais simples impõe o valor de similaridade mínimo que um objeto deve ter para pertencer a um grupo. Logo, quando um objeto é aprovado pela restrição ele é colocado em um grupo.

Os algoritmos básicos de agrupamento são conhecidos por *stars*, *single link*, *strings* e *cliques*. Maiores detalhes sobre o funcionamento destes algoritmos podem ser obtidos em [6].

4. IMPLEMENTAÇÃO

Após haver sido realizado um estudo sobre os principais algoritmos de agrupamento de informações textuais existentes, realizou-se a implementação dos mesmos. Para tanto, adotou-se o sistema operacional *Windows* e a linguagem de programação *C++* sob o ambiente *Borland Cbuilder++*. Além disso, analisando-se seu funcionamento, foram elaborados novos algoritmos (*best-star* e *full-stars*) que nada mais são do que refinamentos com base nos algoritmos existentes, buscando minimizar alguns dos seus problemas.

Até o momento foram realizados alguns experimentos com pequenas coleções de documentos, a fim de verificar se a implementação dos algoritmos esta sendo realizada de forma correta. Porém, mesmo tendo utilizado um número pequeno de documentos, já é possível identificar algumas conclusões que são apresentadas na seção seguinte.

5. CONCLUSÕES

Um dos objetivos do trabalho é validar a ferramenta implementada e os algoritmos elaborados. Para tanto, testes devem ser feitos realizando comparações entre os algoritmos implementados e identificando suas vantagens e deficiências. Alguns testes com pequenas coleções de documentos já foram realizados.

Estes testes indicaram que algoritmos similares ao *Star*, que identificam todos os documentos similares a um único documento, conhecido por *centro da estrela*, são os mais apropriados para utilização em conjunto com técnicas de *Recuperação de Informações* (RI). Isso porque nestes casos é possível garantir que os documentos de um grupo sejam similares ao documento central da estrela, mas não é possível garantir que estes sejam similares entre si. No caso de utilização deste algoritmo em uma ferramenta de RI, o documento central poderia ser o documento indicado pelo usuário em uma operação de busca, enquanto os outros seriam os documentos relevantes retornados.

Além disso, conclui-se que o algoritmo *Cliques* é ideal para gerar grupos cuja finalidade seja a *extração de conhecimento*, já que os grupos identificados possuem documentos altamente similares entre si (devido a

restrição imposta pelo algoritmo). Porém, o algoritmo pode tornar-se muito lento em coleções de documentos muito volumosas, não sendo recomendado para operações de *RI*.

Infelizmente, para que estes resultados possam ser comprovados e validados é necessário utilizar-se de alguma coleção de documentos própria para tal. É possível identificar na literatura uma coleção compilada para fins de comparação – coleção *Reuters-21578* – já conhecida pela comunidade da área.

A coleção *Reuters-21578* foi desenvolvida por *David Lewis* em 1996 para que os sistemas de classificação de informações pudessem ser comparados. Desde então, a coleção passou a ser considerada padrão para testes. A coleção consiste em uma série de artigos (notícias) já categorizados manualmente, e pode ser encontrada em <http://www.research.att.com/~lewis/>. Os resultados de estudos utilizando a coleção podem ser encontrados em publicações como o *ACM-SIGIR* e similares.

De posse da coleção, basta verificar se os grupos de documentos identificados automaticamente são iguais aos grupos produzidos manualmente. Deste modo é possível identificar quais métodos obtêm os melhores resultados, concluindo melhor os estudos. Esta coleção já está disponível para a realização dos testes, faltando somente a realização dos mesmos.

Após realizados os testes e validada a ferramenta e os métodos, outras finalidades que não o agrupamento também são possíveis. É possível pensar em ferramentas que aproveitem os grupos identificados e apliquem sobre eles algum algoritmo de extração de informações (ou conhecimento). Este conhecimento pode ser utilizado em sistemas diversos, capazes de classificar automaticamente informações (é possível pensar em um sistema mais complexo, capaz de analisar o prontuário de um paciente e diagnosticar seu problema). Como exemplo, em trabalho paralelo [7], características peculiares de cada grupo de documentos são coletadas e armazenadas em uma base de conhecimento. Esta base de conhecimento é utilizada em um sistema de recuperação de informações auxiliando o processo de localização de informações.

6. AGRADECIMENTOS

O autor agradece ao apoio financeiro concedido pela agência CAPES, assim como o Instituto de informática da UFRGS, pela utilização de suas dependências, e todo seu pessoal, sempre disposto a cooperar quando possível.

Além disso, o autor gostaria dedicar este trabalho, *em memória*, ao professor e orientador Dr. José Mauro V. de Castilho.

7. REFERÊNCIAS

- [1] L. K. Wives. Um estudo sobre técnicas de recuperação de informações com ênfase em informações textuais. Porto Alegre, Instituto de Informática - UFRGS. Dez, 1997. (Trabalho Individual nº672).
- [2] H. Chen et al. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. MIS Department, University of Arizona, 1996. Disponível por WWW em <http://ai.bpa.arizona.edu/papers/>.
- [3] U. M. Fayyad et al. Advances in Knowledge Discovery and Data Mining. Menlo Park, Cambridge: AAAI, MIT. 1996.
- [4] C. van Rijsbergen. Information Retrieval. Butterworths, London, segunda edição, 1979.
- [5] R. Feldman et al. Exploiting background information in knowledge discovery from text. In: Journal of Intelligent Information Systems. Vol. 9, No.1, Julho/Agosto 1997.
- [6] G. Kowalski. Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers. 1997.
- [7] L. K. Wives; S. Loh. Hyperdictionary: a Knowledge Discovery Tool to Help Information Retrieval. A ser publicado em: Proceedings of the Strings Processing and Information Retrieval - A South American Symposium. IEEE Press. Santa Cruz de La Sierra, Bolívia. Setembro de 1998.