Agrupamento de documentos e o problema da grande dimensionalidade

Thierson Couto Rosa Fernando Almir Nascimento Jr.

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Resumo

A análise de agrupamentos em dados de alta dimensionalidade ainda representa um grande desafio, apesar de todos os avanços que têm sido feito neste tema, em diversas áreas. Em particular, o agrupamento de documentos Web constitui um problema de alta dimensionalidade. Apresentamos um estudo de duas abordagens para o problema da alta dimensionalidade. A primeira delas utiliza o conceito de hipergrafos para mapear as afinidades entre itens de dados. A segunda abordagem estudada utiliza o conceito de vizinho mais próximo compartilhado como métrica de similaridade entre pares de itens.

1 Introdução

A publicação de documentos na Web tem características ímpares que a distingue de todas as demais fontes de documentos. Devido à grande facilidade de composição de documentos multimída, e à liberdade que os autores têm de editar e formatar os seus documentos, a Web se tornou uma fonte muito rica de informações. Por outro lado, a falta de padronização quanto à estrutura dos documentos e a dinamicidade com que eles surgem e são alterados constituem o principal desafio dos mecanismos de pesquisa de documentos na Web.

Tem havido nos últimos anos um esforço no sentido de se desenvolverem técnicas automáticas para organizar e agrupar documentos da Web. O objetivo é o de facilitar o desenvolvimento de sistemas de recuperação de informação de alta qualidade, afetando o mínimo possível a facilidade e liberdade de publicação.

Uma das técnicas utilizadas é o agrupamento automático de documentos, que é o enfoque deste trabalho. As técnicas de agrupamento visam reunir documentos considerados semelhantes entre si em um mesmo grupo, separando em grupos diferentes documentos pouco semelhantes.

O agrupamento de objetos, de um modo geral, é um assunto multidisciplinar e que tem aplicação em diversas áreas. Muitas técnicas de agrupamento publicadas são bastante eficazes e eficientes quando aplicadas a objetos de baixa dimensionalidade. Entretanto, quando o número de atributos dos objetos aumenta em grandes proporções, a maioria das técnicas conhecidas não produz agrupamentos satisfatórios.

Este trabalho enfoca o agrupamento de documentos, que é um problema de alta dimensionalidade. Em particular, estamos interessados em soluções que visam obter bons agrupamentos sem a necessidade de uma redução prévia da dimensionalidade. Nos últimos anos, surgiram diversas propostas nesta direção. Por questão de espaço, descreveremos apenas duas destas propostas. O primeiro trabalho a ser analisado propõe o agrupamento de objetos utilizando regras de associação e hipergrafos [HKKM97]. O segundo, propõe adaptações à técnica tradicional do vizinho mais próximo para o agrupamento em alta dimensionalidade [ESK03].

Os dois trabalhos foram escolhidos por serem elementos representativos de duas abordagens utilizadas para resolver o problema. O primeiro trabalho utiliza regras de associação em

vez de medidas tradicionais de distância. O segundo adapta uma técnica tradicional para que possa ser aplicado a problemas de grande dimensionalidade.

A próxima seção apresenta alguns conceitos básicos sobre agrupamentos de dados em geral. Na segunda seção são apresentadas algumas métricas de distância tradicionais. Na terceira seção, é apresentado o problema da grande dimensionalidade. A quarta seção descreve a utilização de hipergrafos e regras de associação para agrupamentos de dados proposta em [HKKM97]. Na quinta seção é apresentado um algoritmo de agrupamento baseado na técnica do vizinho mais próximo, que pode ser utilizada em aplicações de grande dimensionalidade, proposto em [ESK03]. Na seção 6 são apresentadas as conclusões do trabalho e na seção 7 são apontadas possíveis direções de trabalhos futuros.

2 Medidas de distância

Um algoritmo de agrupamento de objetos necessita de uma medida de semelhança (ou de dessemelhança) entre objetos para decidir se dois objetos *i* e *j* devem pertencer a um mesmo grupo ou não. A medida de semelhança é obtida dos conjuntos de valores de atributos associados aos dois objetos.

Grande parte dos algoritmos tradicionais utiliza o modelo vetorial como base para análise de semelhança entre os objetos. No modelo vetorial, cada objeto é visto como um vetor em um espaço onde as dimensões correspondem aos atributos dos objetos. Assim um objeto pode ser caracterizado pelas coordenadas (valores) nestas dimensões (atributos).

Variações de medidas utilizadas para relacionar vetores em espaço são geralmente utilizadas como medidas de semelhança entre objetos. Um exemplo de medida é a métrica de Minkowski [SEK03] [JMF99], que é uma generalização da distância entre dois pontos no Espaço Euclidiano e é dada pela equação: $P_{ij} = (\Sigma_{k=1}^{d} |x_{ik} - x_{jk}|^{r})^{1/r}$. O valor x_{ik} representa o valor do atributo *k* associado ao objeto *i.* A dimensionalidade é representada por *d* e *r* é um parâmetro de entrada.

No caso do agrupamento de documentos, os termos distintos que ocorrem na coleção correspondem aos atributos. Uma medida de semelhança para documentos bastante utilizada é o cosseno do ângulo formado pelos vetores que representam dois documentos. A medida do cosseno é dada por: $(\Sigma_{i=1}^{t} w_{i,a} \times w_{i,b}) / (\sqrt{\Sigma_{i=1}^{t} w_{i,a}^{2}} \times \sqrt{\Sigma_{i=1}^{t} w_{i,b}^{2}})$, onde *t* corresponde ao número de termos distintos na coleção de documentos, $w_{i,a}$ corresponde a um peso que é associado ao termo *i* em relação ao documento *a* . O peso de cada termo em relação a cada documento é calculado com base na freqüência do termo no documento e na coleção através de uma estratégia de atribuição de pesos denominada *tf-idf.* [YN99].

3 O problema da grande dimensionalidade

Várias técnicas tradicionais de agrupamento são eficazes se a dimensionalidade (número de atributos) dos objetos não é muito grande. Quando o número de dimensões é muito grande, os pontos (objetos) se tornam muito esparsos e as distâncias entre os pontos se tornam relativamente uniformes. Como conseqüência, as métricas se tornam pouco discriminatórias com relação à escolha de quais objetos devem pertencer a um mesmo grupo. Este fenômeno é conhecido como a "maldição da dimensionalidade".

Recentemente, surgiram algumas técnicas para agrupamento de dados de grande dimensionalidade. Algumas destas técnicas são descritas em [SEK03]. Nas próximas seções são discutidas duas dessas técnicas.

4 Agrupamento em Espaço de Grande Dimensionalidade Utilizando Hipergrafos

O algoritmo para agrupamento de dados utilizando hipergrafos foi proposto em [HKKM97]. Ele é constituído de duas fases. Na primeira fase um hipergrafo é construído para representar as relações entre os dados de entrada. Na segunda fase, é aplicado um algoritmo de partição de grafos sobre o hipergrafo gerado. Cada parte do hipergrafo constitui um grupo.

Um hipergrafo $H = (V,E)$ é formado por um conjunto de vértices V e um conjunto de hiperarestas E. Um hipergrafo estende o conceito de grafos permitindo que mais de dois vértices sejam relacionados entre si através de uma hiperaresta. No contexto do algoritmo, o conjunto V corresponde ao conjunto de itens de dados a serem agrupados e uma hiperaresta corresponde a um conjunto itens que têm relação entre si. Os itens de uma hiperaresta são elementos dos conjuntos de itens freqüentes obtidos de regras de associação [AS94] aplicadas aos elementos de V.

4.1 Regras de Associação

As regras de associação foram inicialmente aplicadas para identificar relacionamentos entre produtos presentes nas transações de compras feitas por clientes de uma loja [AS94].

Seja $T=\{t_1,t_2, ...t_n\}$ uma coleção de transações onde cada transação t corresponde a uma lista de itens de dados $t=(i_1, i_2, ... i_k)$, tal que cada item i_j pertence a um conjunto $I = \{ i_1, i_2, ... i_m\}$. Dado um conjunto $C \subseteq I$, define-se o *suporte* de C em relação a T como: $\sigma(C) = |\{t| t \in T, C \subseteq t\}|$. Em outras palavras, $\sigma(C)$ é o numero de transações que contêm o subconjunto de itens C. Por exemplo, dado o conjunto de transações abaixo, $\sigma(\{guardanapo, leite\}) = 3$ e $\sigma = (\{guardanapo, leite, cerveja\}) = 2$.

TID	Itens
1	pão, refrigerante, leite
2	cerveja, pão
3	cerveja, refrigerante, guardanapo, leite
4	cerveja, pão, guardanapo, leite
5	refrigerante, guardanapo, leite

Uma regra de associação é uma expressão do tipo $X \Rightarrow^{s,\alpha} Y$, onde $X \subseteq I$ e $Y \subseteq I$, o suporte s da regra $X \Rightarrow^{s,\alpha} Y$ é definido como $s = \sigma(X \cup Y)/ |T|$ e a confiança α é definida como $\alpha = \sigma(X \cup Y) / \sigma(X)$. A regra {leite, guardanapo} $\Rightarrow$ {cerveja} significa que a presença de leite e guardanapo tende a indicar a presença de cerveja. Esta regra tem suporte $s = \sigma(guardanapo, cerveja, leite) / 5 = 0,4$ e confiança $\alpha = \sigma(guardanapo, cerveja, leite) / \sigma(guardanapo, leite) = 0,66$. O valor alto da confiança em uma regra indica alto grau de correlação entre os conjuntos de itens da regra. O valor alto do suporte indica que a regra existe em um grande número de transações da coleção T.

Um algoritmo para descobrir regras de associação é composto por duas etapas. Na primeira etapa devem ser descobertos todos *conjuntos freqüentes de itens* (subconjuntos de itens que possuem suporte acima de um valor mínimo, fornecido como parâmetro para o algoritmo). A segunda etapa consiste em gerar as regras de associação a partir dos conjuntos freqüentes de itens. Dados dois conjuntos freqüentes {A,B,C,D} e {A,B}, a regra {A,B} $\Rightarrow$ {C,D} será considerada válida se sua confiança $\alpha = \sigma(\{A,B,C,D\})/ \sigma(\{A, B\})$ for superior a um valor mínimo de confiança, fornecido ao algoritmo como parâmetro.

Um dos algoritmos mais conhecidos para se obter os conjuntos grandes de itens é o Apriori. Descreveremos em linhas gerais o funcionamento do algoritmo. Uma descrição completa pode ser encontrada em [AS94].

Inicialmente, o algoritmo faz uma leitura de todas as transações da base de dados e obtém a coleção L_1 com todos os conjuntos freqüentes que possuem um item. Em seguida, obtém uma seqüência L_2, L_3,...L_k de coleções de conjuntos freqüentes de itens. Cada coleção L_j é obtida a partir da coleção anterior L_{j-1}. Este processo é repetido até que uma coleção L_k formada por conjuntos freqüentes com k itens não puder ser obtida.

Em cada passo, o algoritmo considera que os itens dentro de um conjunto estão ordenados lexicograficamente, isto é, dado o conjunto de itens {A, B, C, D}, então A < B < C < D. Uma coleção L_{k+1}, é obtida da coleção L_k,executando-se os passos a seguir:

Passo de junção: para cada par (C_{1k},C_{2k}) de conjuntos freqüentes de itens da coleção L_k, onde $C_{1k} = \{i_1, ...i_k\}$, $C_{2k} = \{j_1, ... j_k\}$, $C_{1k} \neq C_{2k}$ e $i_1 = j_1$, $i_2 = j_2$, ... $i_{k-1} = j_{k-1}$, $i_k < j_k$, obter o novo conjunto $C_{tk+1} = \{i_1, i_2, ... i_{k-1}, i_k, j_k\}$, com k elementos. Incluir C_{tk+1} em uma nova coleção P_{Lk+1}. Por exemplo, Seja L_3 = {{1,2,3},{1,2,4},{1,3,4}, {1,3,5},{2,3,4}}. P_{L4} = {{1,2,3,4}, {1,3,4,5}}. A coleção P_{L4} corresponde a uma versão *preliminar da coleção* L_4.

Passo de poda: para todo conjunto de item $c \in P_{Lk+1}$, retirar c de P_{Lk+1}, se c contém algum subconjunto com k itens que não pertence a L_k . No exemplo, o conjunto c = {1,3,4,5} é retirado de P_{L4}, pois o conjunto de itens {1,4,5} não pertence a L_3. Restará apenas o conjunto {1,2,3,4} em P_{L4}.

Passo de contagem: é verificado se cada conjunto de itens $c \in P_{Lk+1}$ é um conjunto freqüente, isto é, se seu suporte é maior que o suporte mínimo. Os conjuntos que não são freqüentes são eliminados de P_{Lk+1}.

A coleção L_{k+1} corresponde à coleção P_{Lk+1} obtida ao final do passo 3. A saída gerada pelo algoritmo é a seqüência de coleções L_1, L_2, ... L_{k+1} .

4.2 Modelagem de itens de dados por hipergrafos

Um problema importante na modelagem de itens de dados por um hipergrafo é o de determinar os itens de dados que possam ser agrupados em hiperarestas. Outro problema, é a atribuição de pesos às hiperarestas.

Em [HKKM97] os conjuntos freqüentes de itens de dados relacionados a regras de associação são utilizados como hiperarestas do hipergrafo. Os conjuntos freqüentes de itens de dados são considerados muito bons candidatos a se tornarem hiperarestas pelo fato de capturarem relacionamentos entre mais de dois itens de dados.

Em [HKKM97], os pesos das hiperarestas foram calculados com base nas confianças das regras de associação relacionadas aos conjuntos freqüentes. Para cada hiperaresta, foram consideradas apenas as regras de associação que envolviam todos os itens da hiperaresta e que possuíam apenas conjuntos unitários à direita do operador $\Rightarrow$.

4.3 Agrupamento de itens de dados utilizando hipergrafos

Os conjuntos freqüentes de itens de dados por si só já representam relacionamentos entre itens de uma coleção de transações. Mas alguns relacionamentos indiretos não são captados utilizando-se apenas os conjuntos mais freqüentes. Por exemplo, sejam os seguintes conjuntos freqüentes de itens obtidos de uma coleção de transações de vendas: {fralda, talco, leite em pó}, {mamadeira, fralda} e {mamadeira, leite em pó}. Intuitivamente estes produtos estão todos

relacionados. Porém um conjunto contendo os itens: mamadeira, fralda, leite em pó e talco pode não ser um conjunto freqüente de itens. Logo, não seria produzido pelo algoritmo Apriori.

Um hipergrafo pode ser utilizado para reunir conjuntos freqüentes de itens de dados em grupos maiores. Dado que as hiperarestas já representam os conjuntos freqüentes de itens, o agrupamento maior pode ser conseguido através de uma partição do hipergrafo. Como as hiperarestas representam relacionamentos intraconjuntos de itens, uma partição deve minimizar o peso das hiperarestas que são cortadas pela partição.

A qualidade da partição do hipergrafo pode ser avaliada segundo um método proposto em [HKKM97]. O método de avaliação mede a razão entre o peso de todas as hiperarestas dentro do grupo *G* e o peso de todas arestas que têm pelo menos um vértice em *G*. Se o valor da função for inferior a um valor mínimo, o grupo é descartado.

O problema de obter uma bisseção ótima de um grafo ou hipergrafo é NP-completo. A heurística de partição de hipergrafos utilizados nos experimentos realizados em [HKKM97] foi a HMETIS [KAKS97]. Este algoritmo gera uma bisseção do hipergrafo original. Para obter mais partições do hipergrafo, o algoritmo deve ser executado repetidas vezes, a cada vez é gerada uma bisseção de uma das metades originadas em execuções anteriores.

HMETIS é uma heurística de vários níveis para partição de hipergrafos. Em um algoritmo de vários níveis existem três fases: a *fase de engrossamento*, a *fase de partição* e a *fase de refinamento*. A fase de engrossamento é repetida várias vezes. A cada iteração um hipergrafo menor é obtido utilizando uma técnica de fusão de vértices [KAKS97], [KK98]. O número de vezes que a fase é executada depende de caracteríticas de conectividade do grafo e pode ser controlado por parâmetro.

Na fase de partição, o algoritmo permite a escolha de várias técnicas de bisseção (bisseção espectral, algoritmo Kernighan-Lin, ou algoritmo de crescimento – mais detalhes podem ser encontrados em [KK98]). O objetivo de todas estas técnicas é o de gerar uma bisseção no hipergrafo de tal modo a minimizar o corte de hiperarestas.

A fase de refinamento é repetida o mesmo número de vezes que a fase de engrossamento. A cada vez são desfeitas as fusões de vértices feitas durante a fase de engrossamento e são retomadas as hiperarestas que haviam desaparecido. Com a inserção destes vértices e hiperaretas a bisseção deve ser adaptada para que o menor número de hiperarestas sejam cortadas. Feita esta adaptação, uma nova repetição do refinamento é iniciada. O processo termina quando um hipergrafo com o mesmo número de vértices do hipergrafo original for obtido.

4.4 Análise

A complexidade de tempo do algoritmo Apriori é linear em relação ao número de transações [HKKM97]. Entretanto, naquele trabalho é relatado que o Apriori apresenta bom desempenho quando o suporte mínimo é considerado grande. Caso contrário, muito conjuntos de itens são gerados, degradando sensivelmente o tempo de execução. Entretanto, na maioria das aplicações o suporte mínimo deve ser grande, para que os conjuntos freqüentes sejam significativos.

A coplexidade de tempo do HMETIS para gerar partições com k grupos é $O((V+E) \log k)$, onde V é o conjunto de vértices (itens de dados) e E é o número de hiperarestas.

4.5 Aplicação de hipergrafos a agrupamentos de textos

Um algoritmo para gerar conjuntos freqüentes de itens pode ser utilizado para gerar conjuntos de textos de uma coleção de documentos. Neste caso, cada termo do índice invertido corresponde a

uma trasação. Os conjuntos freqüentes de documentos gerados se tornam hiperarestas de um grafo e um algoritmo de partição de hipergrafo pode ser utilizado para gerar os grupos. Em [HKKM97] é relatado um experimento em que foi utilizada uma versão adaptada do Apriori, o Min-Apriori, e o HMETIS.

5 O algoritmo do vizinho mais próximo compartilhado

No problema do agrupamento de documentos, o que distingue os documentos de um mesmo tema é a freqüência com que as palavras são usadas, sendo que cada tema possui um vocabulário típico de palavras. Por exemplo, documentos sobre futebol geralmente falam de times, partidas, jogadores, etc., enquanto documentos sobre política falam de partidos, votações, eleitores, etc. No entanto, os vocabulários específicos podem se sobrepor, documentos podem conter palavras de diferentes vocabulários e um documento pode conter palavras de um determinado vocabulário mesmo se ele não estiver relacionado ao tema que em geral usa esse vocabulário.

Em algumas coleções TREC, por exemplo, de 15 a 20 por cento dos vizinhos mais próximos de um documento são de classes ou temas diferentes, quando é utilizada uma métrica de distância ou similaridade direta entre dois documentos [SEK03]. Isso torna necessária a definição de uma métrica de distância ou similaridade que seja mais confiável em altas dimensões.

O método que vamos apresentar nesta seção foi proposto em [ESK03] e se baseia em um conceito conhecido como **vizinho mais próximo compartilhado**. Vamos utilizar a terminologia de grafos para explicar esse método.

Inicialmente, são encontrados os *k* vizinhos mais próximos de cada um dos pontos, com base em uma métrica qualquer de similaridade direta, como a distância cosseno, por exemplo. É então criado um grafo (como o da Figura 1a, por exemplo) que iremos chamar de **grafo do vizinho mais próximo**, no qual irá existir uma aresta entre dois pontos se e somente se um deles fizer parte da lista dos *k* vizinhos mais próximos do outro, e vice-versa. O peso de uma aresta que conecta dois pontos será dado pela medida da distância cosseno entre esses dois pontos.

É definido então um novo grafo, que iremos chamar de **grafo do vizinho mais próximo compartilhado**, em que existirá uma aresta entre dois pontos se e somente se houver uma aresta entre os mesmos dois pontos no grafo do vizinho mais próximo, e o peso da aresta será o número de vizinhos mais próximos compartilhados entre os dois pontos (Figura 1b). Esse valor utilizado como peso das arestas passa a ser então a nova métrica de similaridade entre dois pontos, que serão ditos similares se compartilharem um número mínimo de vizinhos mais próximos.

O próximo passo é calcular a densidade em cada ponto. Vamos definir a densidade em um ponto como sendo o número de pontos vizinhos a ele cuja similaridade seja maior ou igual a um valor especificado pelo usuário, que chamaremos de *Eps*. No grafo da Figura 1c, por exemplo, o valor utilizado de *Eps* foi 2. Ou seja, para calcular a densidade em cada ponto, contamos o número de vizinhos desse ponto cuja similaridade com ele era maior ou igual a 2. Alternativamente, poderíamos definir a densidade em um ponto como sendo a soma dos pesos de todas as arestas que se conectam a ele.

São encontrados então os chamados “pontos centrais”, que correspondem aos pontos cuja densidade seja maior ou igual a um determinado valor, também especificado pelo usuário, ao qual chamaremos de *MinPts*. No grafo da Figura 2a, utilizamos um valor de *MinPts* igual a 5, e apenas os pontos com densidade maior ou igual a 5 foram marcados como pontos centrais. Além disso, todos os pontos centrais cuja similaridade par a par era maior ou igual a *Eps* (que no nosso exemplo vale 2) foram colocados no mesmo grupo.

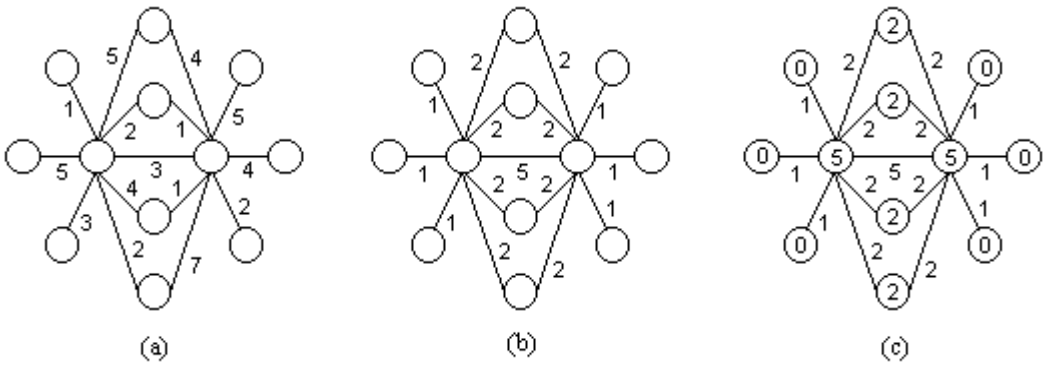

Figura 1: (a) grafo do vizinho mais próximo (b) grafo do vizinho mais próximo compartilhado (c) densidade no grafo do vizinho mais próximo compartilhado.

No próximo passo, todos os pontos não-centrais que não forem similares a nenhum ponto central por uma medida maior ou igual a *Eps* são considerados ruído e descartados. No nosso exemplo, como o valor de *Eps* é 2, pudemos descartar todos os pontos cuja máxima similaridade com um ponto central era 1 (Figura 2b).

Por fim, os pontos não-centrais que restarem são agrupados com o ponto central que for mais similar. No nosso exemplo, foi formado um único grupo contendo 6 pontos (Figura 2c).

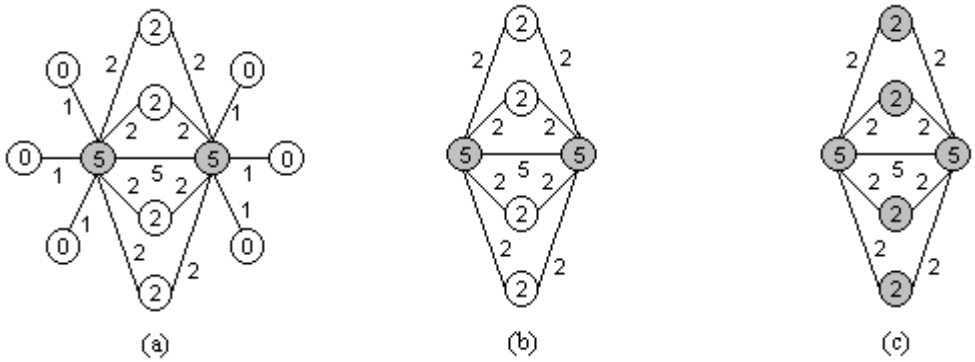

Figura 2: (a) pontos centrais no grafo do vizinho mais próximo compartilhado (b) eliminação do ruído (c) formação dos grupos

Como pudemos notar através do nosso exemplo, o método do vizinho mais próximo compartilhado não agrupa todos os pontos. De fato, ele elimina todos os pontos que parecem ser ruído. Caso se deseje agrupar todos os pontos, é necessário um passo adicional, no qual cada ponto não-agrupado é associado ao grupo que contenha o ponto central mais similar a ele.

Os parâmetros utilizados no algoritmo, *Eps* e *MinPts*, além do tamanho *k* da lista de vizinhos mais próximos de um ponto, podem ser variados para se obter uma maior resolução ou granularidade do agrupamento e para permitir que o usuário controle quantos pontos serão agrupados e quantos serão considerados ruído.

5.1 Análise

A utilização do conceito do vizinho mais próximo compartilhado permite que o método apresentado consiga tratar de maneira mais eficaz o problema da variação de densidade e o problema da alta dimensionalidade. A utilização do conceito de pontos centrais permite que grupos de diferentes formas e tamanhos sejam tratados de forma adequada.

O método do vizinho mais próximo compartilhado possui uma complexidade de tempo da ordem de $O(n^2)$, onde n é o número de pontos, se as similaridades entre os pares de pontos tiverem que ser calculadas para se encontrarem os k vizinhos mais próximos.

Para melhorar a eficiência do processo de criação das listas de vizinhos mais próximos, algumas técnicas podem ser utilizadas. No caso específico do agrupamento de documentos, a eficiência pode ser melhorada a partir da utilização de índices invertidos, como demonstrado em [ESK03].

6 Conclusões

Neste trabalho, foram apresentados alguns métodos de agrupamento de documentos que indicam fornecer melhores resultados que os métodos de agrupamento tradicionais quando a dimensionalidade dos dados é muito alta. Os métodos apresentados foram o agrupamento baseado em modelos de hipergrafo [HKKM97] e o método do vizinho mais próximo compartilhado [ESK03].

O método de agrupamento baseado em modelos de hipergrafo mapeia os relacionamentos presentes nos dados usando um modelo de hipergrafo. As hiperarestas do hipergrafo representam as afinidades entre subconjuntos dos dados e os pesos das hiperarestas representam as intensidades dessas afinidades. As afinidades entre os itens são obtidas usando-se o conceito de regras de associação [AS94]. É usado então um algoritmo de particionamento para gerar os grupos.

Os resultados reportados em [HKKM97] mostram que o algoritmo de agrupamento baseado em modelos de hipergrafo é mais eficaz que os métodos tradicionais, mesmo não exigindo uma redução da dimensionalidade. Além disso, o algoritmo permite que a qualidade dos grupos encontrados seja variada de acordo com os requisitos exigidos. Entretanto, o algoritmo requer bons parâmetros para produzir bons resultados. Além disso o algoritmo de particionamento de hipergrafos utilizado apresenta algumas deficiências, não sendo capaz de identificar automaticamente o número ideal de partições e, muitas vezes, “cortando” hiperarestas que representam relacionamentos significativos.

O método do vizinho mais próximo compartilhado propõe uma métrica de similaridade mais “confiável” que as métricas de similaridade tradicionais, uma vez que estas, em geral, tendem a perder a eficácia quando a dimensionalidade dos dados aumenta muito. No algoritmo apresentado, a similaridade entre dois pontos é obtida usando-se uma métrica tradicional e é “confirmada” pelo número de vizinhos mais próximos compartilhados entre esses dois pontos. Adicionalmente, o algoritmo utiliza o conceito de “pontos centrais” para tratar de grupos com diferentes formas e tamanhos.

O ponto forte do método do vizinho mais próximo compartilhado é que ele consegue encontrar grupos de baixa e média densidade em regiões uniformes, onde outros métodos tendem a não encontrar nada. Entretando, o algoritmo não agrupa todos os pontos, o que às vezes é necessário, além de exigir uma interação muito grande do usuário até que sejam encontrados os parâmetros ideais. Além disso, o algoritmo apresenta uma complexidade de tempo quadrática, embora algumas otimizações tenham sido propostas.

7 Trabalhos futuros

Os dois métodos apresentados foram avaliados em condições muito específicas, o que não poderia ter sido diferente. Porém, tais métodos devem ser avaliados também em outros domínios, onde diferentes condições poderiam exigir novas abordagens ou mesmo invalidar a utilização de tais métodos em um domínio específico.

No caso particular do agrupamento de documentos Web, esses métodos foram testados em subconjuntos de documentos que não necessariamente representam de maneira fiel a Web como um todo. Um trabalho que poderia ser feito é a avaliação de um desses métodos aplicado exclusivamente no problema do agrupamento de documentos Web, onde poderiam ser utilizadas amostras de documentos que fossem as mais representativas possível, e onde o método poderia ser complementado com abordagens que melhorassem a sua eficácia neste caso específico.

No método de agrupamento usando modelos de hipergrafos, podem ser avaliados algoritmos de particionamento de hipergrafos que sejam mais adequados para diferentes tipos de dados, uma vez que o algoritmo usado no trabalho (HMETIS) possui algumas limitações. Novamente, isso poderia ser feito com ênfase no problema de agrupamento de documentos Web, procurando otimizar ao máximo o emprego dos modelos de hipergrafo neste domínio em especial.

Finalmente, a utilização de regras de associação se mostrou uma forma eficaz e intuitiva para encontrar afinidades entre itens, sugerindo que esta técnica possa ser utilizada em combinação com outras técnicas para obter resultados diferentes dos obtidos com modelos de hipergrafos.

Referências

[AS94] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, In Proceedings of 20th VLDB Conference, pp. 487-499, Santiago, Chile, 1994.

[ESK03] Levent Ertoz, Michael Steinbach, and Vipin Kumar, “Finding Clusters of Different Sizes, Shapes and Densities in Noisy, High Dimensional Data”, SIAM International Conference on Data Mining, San Francisco, CA, May, 2003.

[HKKM97] Eui-Hong Han, George Karypis, Vipin Kumar, and Bomshad Mobasher, “Clustering In High-Dimensional Sapace Using Hypergraph Models “, Techinical Report TR-97-063, Department of Computer Science, University of Minesota, Minneapolis, Minesota, 1997.

[JMF99] A. K. Jain, M. N. Murty and P. J.Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol 31, No. 3, September, 1999.

[KAKS97] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel Hipergraph Partitioning: Applications in VLSI Design", Proc. ACM/IEEE Design Automation Conference, pp. 526-529,1997.

[KK98] G. Karpis, and V. Kumar, “A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs, SIM Journal of Scientific Computing, Vol. 20, 1998.

[SEK03] Michael Stinbach, Levent Ertöz and Vipin Kumar, “The Chalenges of Clustering High Dimensional Data”, New Vistas in Statistical Physics -- Applications in Econophysics, Bioinformatics, and Pattern Recognition, forthcoming 2003, Springer-Verlag.

[YN99] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, “Modern Information Retrieval”, ACM Press, 1999.