UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

JULIANA BONATO DOS SANTOS

Automatizando o Processo de Estimativa de Revocação e Precisão de Funções de Similaridade

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. Carlos Alberto Heuser
Orientador

Profª. Drª. Viviane Moreira Orengo
Co-orientadora

Porto Alegre, setembro de 2008.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Santos, Juliana Bonato dos

Automatizando o Processo de Estimativa de Revocação e Precisão de Funções de Similaridade / Juliana Bonato dos Santos – Porto Alegre: Programa de Pós-Graduação em Computação, 2008.

61 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2008. Orientador: Carlos Alberto Heuser; Co-orientadora: Viviane Moreira Orengo.

1.Validação do processo de agrupamento. 2.Agrupamento por similaridade. 3.Funções de similaridade. 4.Revocação e precisão. I. Heuser, Carlos Alberto. II. Orengo, Viviane Moreira. III. Título.

“Aos meus pais Joel Rosa dos Santos e Denise Bonato dos Santos.”

AGRADECIMENTOS

Agradeço imensamente aos meus pais, Joel e Denise, por todo o apoio, incentivo, carinho e compreensão.

À minha família, em especial a minha avó, por toda a sua dedicação ao longo da minha vida.

Ao meu orientador, Prof. Dr. Carlos A. Heuser, e a minha co-orientadora, Prof[a]. Dr[a]. Viviane M. Orengo, pela orientação e imenso apoio durante o desenvolvimento deste trabalho.

Aos meus queridos amigos, por todos os momentos de alegria e descontração que me proporcionam.

Aos meus colegas do TRF da 4[a] Região, por todo apoio e incentivo.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

BD	Banco de Dados
ESS	*Error Sum of Squares*
FEBRL	*Freely Extensible Biomedical Record Linkage*
FERP	Ferramenta para Estimativa de Revocação e Precisão
HACM	*Hierarchical Agglomerative Clustering Methods*
MAP	*Mean Average Precision*
R&P	Revocação e Precisão
RI	Recuperação de Informações

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Os mecanismos tradicionais de consulta a bases de dados, que utilizam o critério de igualdade, têm se tornado ineficazes quando os dados armazenados possuem variações tanto ortográficas quanto de formato. Nesses casos, torna-se necessário o uso de funções de similaridade ao invés dos operadores booleanos. Os mecanismos de consulta por similaridade retornam um *ranking* de elementos ordenados pelo seu valor de similaridade em relação ao objeto consultado. Para delimitar os elementos desse *ranking* que efetivamente fazem parte do resultado pode-se utilizar um limiar de similaridade. Entretanto, a definição do limiar de similaridade adequado é complexa, visto que este valor varia de acordo com a função de similaridade usada e a semântica dos dados consultados. Uma das formas de auxiliar na definição do limiar adequado é avaliar a qualidade do resultado de consultas que utilizam funções de similaridade para diferentes limiares sobre uma amostra da coleção de dados.

Este trabalho apresenta um método automático de estimativa da qualidade de funções de similaridade através de medidas de revocação e precisão computadas para diferentes limiares. Os resultados obtidos a partir da aplicação desse método podem ser utilizados como metadados e, a partir dos requisitos de uma aplicação específica, auxiliar na definição do limiar mais adequado. Este processo automático utiliza métodos de agrupamento por similaridade, bem como medidas para validar os grupos formados por esses métodos, para eliminar a intervenção humana durante a estimativa de valores de revocação e precisão.

Palavras-Chave: validação do processo de agrupamento, agrupamento por similaridade, funções de similaridade, revocação e precisão.

Automatizing the process of estimating recall and precision of similarity functions

ABSTRACT

Traditional database query mechanisms, which use the equality criterion, have become inefficient when the stored data have spelling and format variations. In such cases, it's necessary to use similarity functions instead of boolean operators. Query mechanisms that use similarity functions return a ranking of elements ordered by their score in relation to the query object. To define the relevant elements that must be returned in this ranking, a threshold value can be used. However, the definition of the appropriated threshold value is complex, because it depends on the similarity function used and the semantics of the queried data. One way to help to choose an appropriate threshold is to evaluate the quality of similarity functions results using different thresholds values on a database sample.

This work presents an automatic method to estimate the quality of similarity functions through recall and precision measures computed for different thresholds. The results obtained by this method can be used as metadata and, through the requirements of an specific application, assist in setting the appropriated threshold value. This process uses clustering methods and cluster validity measures to eliminate human intervention during the process of estimating recall and precision.

Keywords: cluster validity, clustering, similarity functions, recall and precision.

1 INTRODUÇÃO

Quando pessoas diferentes inserem dados em um sistema, grandes variações podem ocorrer devido a erros de ortografia, erros de acentuação, utilização de diferentes formatos de dados, uso de abreviaturas ou siglas, caracteres trocados, entre outros. O mesmo problema ocorre em sistemas que foram integrados, em que dados provenientes de diferentes bases de dados podem estar representados de várias formas embora façam referência ao mesmo objeto do mundo real. Nesses casos, os processadores tradicionais de consulta (ULLMAN; GARCIA-MOLINA; WIDOM, 2002) não possuem suporte quando as consultas envolvem tais variações, visto que se baseiam na busca pelo critério de igualdade.

Por exemplo, em uma base de dados de referências bibliográficas, que permite ao usuário digitar o nome da conferência em que tenha publicado um determinado artigo, podemos encontrar uma mesma conferência cadastrada em alguns artigos como "*VLDB*" e em outros como "*Very Large Database*". Embora esses nomes estejam escritos de formas diferentes, ambos representam à mesma conferência. Entretanto, ao tentarmos recuperar todos os artigos publicados em uma determinada conferência, utilizando como predicado da consulta o nome "*VLDB*", por exemplo, e mecanismos de consulta que utilizam critérios de igualdade, nem todos os artigos correspondentes seriam retornados. Portanto, nesse caso, para recuperar todas as ocorrências da conferência citada, o critério de igualdade não se aplica, tornando necessária a utilização de um mecanismo de consulta que suporte a busca por similaridade.

Uma função de similaridade recebe dois elementos como entrada e retorna um valor de similaridade, também conhecido como escore, entre 0 e 1. Assim, as consultas por similaridade retornam como resultado um *ranking* de elementos, isto é, uma lista contendo todos os elementos da base de dados ordenados de acordo com o seu escore em relação ao objeto de consulta. As funções de similaridade têm sido amplamente utilizadas por diversas aplicações tais como: junção por similaridade (COHEN, 2000), deduplicação de registros (WINKLER, 1999; FELLEGI; SUNTER, 1969), *entity resolution* (BENJELLOUN et al., 2006), consultas imprecisas (NAMBIAR; KAMBHAMPATI, 2003), entre outras. Nessas aplicações, as funções de similaridade são utilizadas para identificar se duas instâncias (cadeias de caracteres, registros, documentos XML, etc.) podem ser consideradas como sendo o mesmo objeto do mundo real. Neste contexto, é importante estabelecer uma estratégia para separar os elementos que efetivamente representam o objeto de consulta, ou seja, que são relevantes, dos demais.

Uma das estratégias apresentadas na literatura para delimitar os elementos retornados por consultas por similaridade é a definição de um valor mínimo de escore aceitável (também conhecido como limiar ou *threshold*) (SCHALLEHN; SATTLER;

SAAKE, 2004; ORTEGA-BINDERBERGER, 2002; MOTRO, 1988). Assim, os elementos cujo escore de similaridade em relação ao objeto de consulta seja igual ou superior a esse limiar são definidos como parte do resultado. Dessa forma, esse valor implica diretamente na qualidade dos resultados dessas consultas. Caso seja definido um limiar muito baixo, podem ser retornados elementos que não são relevantes, ou seja, "*falsos positivos*", caso contrário, elementos relevantes podem não ser retornados, ou seja, "*falsos negativos*". No contexto deste trabalho, utiliza-se o termo consulta por abrangência para designar as consultas por similaridade que possuem um limiar de escore aceitável.

Para determinar o limiar adequado a ser usado nas consultas por abrangência é necessário avaliar vários *rankings* gerados com diferentes limiares. Assim, a partir da identificação, por parte de um especialista humano, de quais elementos são relevantes à consulta, é possível identificar o *ranking* que gera o melhor resultado. Nesse caso, a forte dependência da intervenção de um especialista humano dificulta a avaliação de consultas por abrangência na prática.

Foi proposto por Stasiu (STASIU; HEUSER; SILVA, 2005; STASIU, 2007), um método semi-automático de estimativa de qualidade de consultas por abrangência para vários limiares, utilizando amostras de base de dados, através de medidas de revocação (*recall*) e precisão (*precision*) (R&P) (BAEZA-YATES; RIBEIRO-NETO, 1999). Nesse método, ao invés de identificar os elementos que são relevantes para cada consulta, o especialista humano apenas informa o número de objetos do mundo real que estão representados em cada amostra. A partir dessa informação, um algoritmo de agrupamento por similaridade (*clustering*) gera grupos de elementos. Cada grupo gerado deve conter apenas os elementos que representam um único objeto do mundo real. Esses grupos são utilizados para identificar quais são os elementos relevantes quando um determinado elemento é utilizado como objeto de consulta, isto é, todos os elementos pertencentes ao grupo do objeto de consulta são considerados relevantes à consulta. Uma limitação desse método é que o tamanho da amostra não deve ser muito grande para permitir a identificação, pelo especialista humano, do número de objetos distintos.

Como resultado, este método semi-automático gera uma tabela de valores estimados de R&P para 11 diferentes limiares (0.0, 0.1, 0.2, ..., 1.0). A partir dessa tabela e levando em conta alguns requisitos específicos para cada aplicação, tais como precisão alta, revocação alta ou a melhor combinação entre ambas as medidas, é possível determinar qual o limiar mais adequado a ser utilizado.

O presente trabalho visa automatizar este método de estimativa de valores de R&P, eliminando completamente a intervenção do especialista humano, através da utilização de medidas de qualidade do resultado de algoritmos de agrupamento por similaridade. Conforme demonstrado pelos experimentos apresentados no Capítulo 4, quando o valor de um coeficiente que mede a qualidade do resultado retornado por um processo de agrupamento por similaridade, denominado *silhouette coefficient* é maximizado (TAN; STEINBACH; KUMAR, 2006; ARANGANAYAGI; THANGAVEL, 2007), cada grupo tende a conter instâncias de dados que são representações de um único objeto do mundo real. Assim, este agrupamento é escolhido como o mais adequado para o conjunto de dados em análise. Deve-se observar que o *silhouette coefficient* é uma medida baseada em critérios internos (MANNING; RAGHAVAN; SCHÜTZE, 2008), isto é, não necessita que os elementos do conjunto de dados estejam previamente

avaliados. Dessa forma, torna-se possível realizar todo o processo de estimativa de valores de R&P sem a necessidade da intervenção de um especialista humano.

1.1 Contribuições

A principal contribuição deste trabalho é automatizar o processo de estimativa de R&P de funções de similaridade através de uma medida, baseada em critérios internos, de qualidade dos grupos gerados por algoritmos de agrupamento por similaridade.

Como contribuições secundárias podem ser citadas:

- Estudo e análise de medidas de validação do processo de agrupamento por similaridade;
- Avaliação da correlação dos valores obtidos com uma medida de validação baseada em critérios internos (*silhouette coefficient*) e uma medida de validação baseada em critérios externos (medida F);
- Avaliação da semelhança entre os valores estimados de R&P de treinamento e valores de R&P de teste; e
- Avaliação da semelhança entre os valores estimados de R&P obtidos com a aplicação do método automático e os valores estimados de R&P obtidos com a aplicação do método semi-automático.

1.2 Organização do Texto

Esta dissertação está estruturada da seguinte forma:

- Capítulo 2, Agrupamento por similaridade na estimativa de valores de R&P, apresenta os principais conceitos relacionados aos métodos de agrupamento por similaridade, bem como medidas de validação dos agrupamentos gerados por esses métodos. Além disso, apresenta o método clássico e o método semi-automático de estimativa de valores de R&P para funções de similaridade;
- Capítulo 3, Método automático de estimativa de valores de R&P, descreve o método automático de estimativa de R&P para funções de similaridade;
- Capítulo 4, Avaliação experimental, apresenta a forma de validação dos experimentos realizados, assim como os procedimentos realizados e os resultados obtidos;
- Capítulo 5, Conclusão, apresenta as conclusões e os possíveis trabalhos futuros.

2 AGRUPAMENTO POR SIMILARIDADE NA ESTIMATIVA DE VALORES DE R&P

Neste capítulo, são apresentados os principais conceitos relacionados aos métodos de agrupamento por similaridade, bem como algumas medidas encontradas na literatura para validação dos agrupamentos gerados por esses métodos. Esses conceitos são importantes, visto que o método automático de estimativa de valores de R&P está baseado na identificação do agrupamento que melhor particiona o conjunto de dados em análise. Além disso, são apresentados, também, o método clássico e o método semi-automático de estimativa de valores de R&P.

A Seção 2.1 inicia apresentando alguns conceitos sobre o processo de agrupamento por similaridade e, a seguir, apresenta os principais algoritmos de agrupamento que foram utilizados nos experimentos. Na Seção 2.2, são descritos os principais métodos de validação desse processo de agrupamento. A Seção 2.3 descreve as etapas correspondentes aos métodos clássico e semi-automático de estimativa de valores de R&P.

2.1 Processo de Agrupamento por Similaridade

Nesta seção são apresentados, de forma resumida, os principais conceitos relacionados ao processo de agrupamento por similaridade. Além disso, são descritos os principais algoritmos de agrupamento que foram utilizados nos experimentos realizados para validação do método automático de estimativa de R&P. Um estudo mais detalhado sobre processos de agrupamento pode ser encontrado nos trabalhos desenvolvidos por Hartingan (HARTIGAN, 1975), Aldenderfer (ALDENDERFER; BLASHFIELD, 1984), Jain (JAIN; MURTY; FLYNN, 1999), Everitt (EVERITT; LANDAU; LEESE, 2001), Wives (WIVES, 2004) e Manning (MANNING; RAGHAVAN; SCHÜTZE, 2008), entre outros.

Os algoritmos de agrupamento (*clustering*) visam organizar os elementos em grupos (*clusters*) de elementos semelhantes (ALDENDERFER; BLASHFIELD, 1984). Neste trabalho, são utilizadas funções de similaridade para identificar o grau de semelhança entre os elementos. Os grupos formados devem apresentar alto grau de similaridade entre seus elementos internos e baixo grau de similaridade em relação aos elementos pertencentes a outros grupos (JAIN; MURTY; FLYNN, 1999, MANNING; RAGHAVAN; SCHÜTZE, 2008). Assim, cada grupo deve conter todos os elementos que representam um único objeto do mundo real.

No contexto deste trabalho, o processo de agrupamento por similaridade é utilizado para estimar os valores de R&P sem a necessidade de intervenção de um especialista humano. Informações utilizadas para o cálculo das medidas de revocação e precisão, tais como o número de elementos relevantes em cada consulta por abrangência, podem ser extraídas dos grupos gerados nesse processo.

Segundo Hartingan (1975), quanto à sua disposição, os grupos podem ser hierárquicos ou não-hierárquicos. Os grupos hierárquicos são agrupamentos que possuem a estrutura de uma árvore, possuindo, portanto, uma relação de hierarquia entre os grupos formados. Já os agrupamentos não-hierárquicos são agrupamentos que se constituem de partições isoladas ou disjuntas, sem nenhuma relação de hierarquia (CUTTING et al., 1992). A Figura 2.1 apresenta exemplos de um agrupamento hierárquico e um agrupamento não-hierárquico.

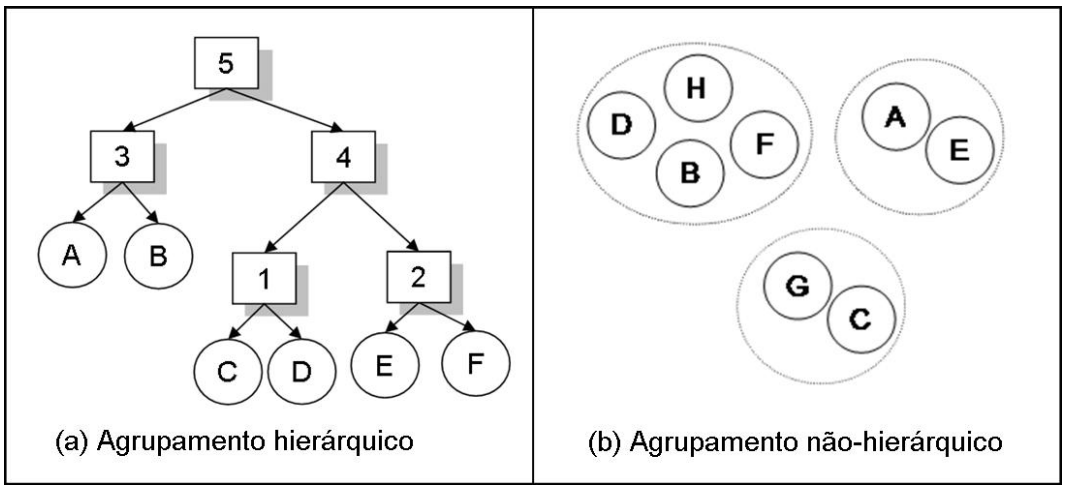


Figura 2.1: Exemplo de um agrupamento hierárquico *versus* um não-hierárquico.

No trabalho desenvolvido por Jain (JAIN; MURTY; FLYNN, 1999) é apresentada uma taxonomia das diferentes abordagens para o processo de agrupamento. Essa taxonomia está dividida em dois grandes grupos: métodos hierárquicos (*hierarchical clustering*) e métodos de particionamento (*partitional clustering*).

Já Aldenderfer (ALDENDERFER; BLASHFIELD, 1984) classifica os métodos de agrupamento nas seguintes categorias: hierárquicos aglomerativos (*hierarchical aglomerative*), hierárquicos divisivos (*hierarchical divisive*), de particionamento iterativo (*iterative partitioning*), de busca em profundidade (*density search*), fator-analítico (*factor-analytic*), de amontoamento (*clumping*) e grafo-teóricos (*graph-theoretic*). A seguir, são apresentadas, de forma resumida, as três primeiras categorias citadas acima.

2.1.1 Métodos Hierárquicos Aglomerativos

Os métodos hierárquicos aglomerativos (conhecidos na literatura pelo acrônimo HACM – *Hierarchical Agglomerative Clustering Methods*) possuem uma abordagem "*bottom-up*", ou seja, iniciam seu processamento considerando que cada elemento corresponde a um grupo e então unem pares de grupos sucessivamente até que todos os elementos tenham sido unidos em um único grupo (MANNING; RAGHAVAN;

SCHÜTZE, 2008). Como resultado, esses métodos produzem uma estrutura em árvore binária, referenciada na literatura como dendograma. Um exemplo de um dendograma pode ser visto na Figura 2.2. Nesse exemplo, o eixo *y* representa o valor de similaridade com o qual os nodos (grupos) foram unidos. Os nodos folhas do dendograma (as letras A, B, C, D, E, F e G da Figura 2.2) correspondem aos elementos do conjunto de dados que está sendo agrupado.

Para transformar a estrutura hierárquica do dendograma em grupos cada nodo interno é comparado contra um limiar de similaridade. Essa comparação é realizada de forma "*top-down*", ou seja, a análise inicia a partir do nodo pai. Caso o valor de similaridade do nodo em análise for maior ou igual ao limiar, todos os nodos folhas que se encontram na subárvore desse nodo são unidos em um único grupo. Caso contrário, o processo prossegue analisando seus nodos filhos. Assim, o número de grupos resultantes varia de acordo o limiar de similaridade estabelecido para o processo de agrupamento.

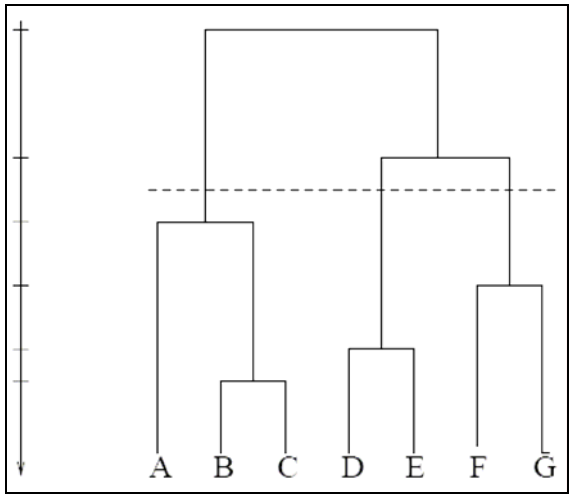


Figura 2.2: Ilustração de um dendograma (JAIN; MURTY; FLYNN, 1999)

Na Figura 2.3 podemos visualizar um exemplo dos grupos que seriam formados, a partir de um dendograma, caso fossem utilizados limiares iguais a 0.6 e 0.8.

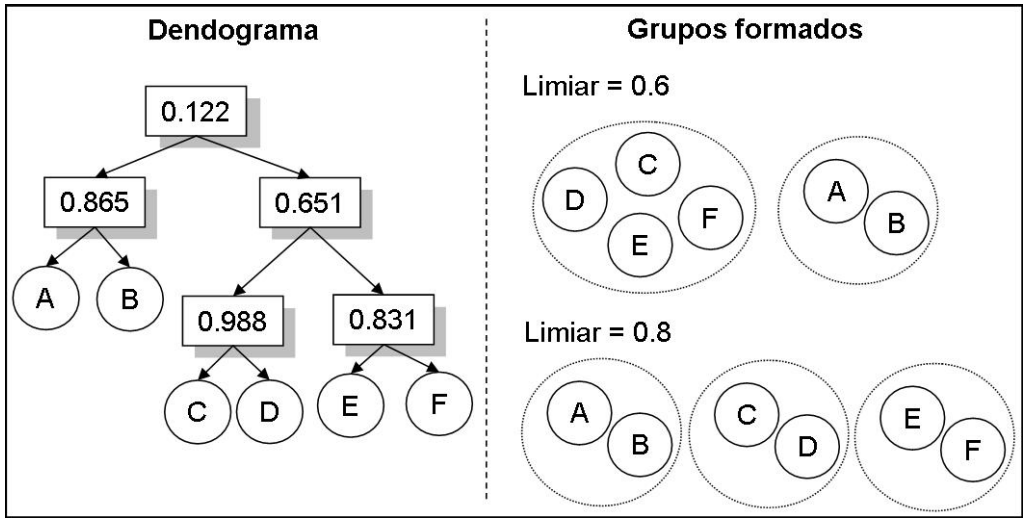


Figura 2.3: Exemplo da geração de grupos a partir de um dendograma

Os métodos hierárquicos aglomerativos aninham os objetos destacando as relações de composição, abrangência e especificidade entre os elementos. Além disso, os algoritmos que implementam esses métodos tendem a ser rápidos, necessitando de $n - 1$ passos para gerar o dendograma, sendo que n corresponde ao número de elementos do conjunto de dados. A maior desvantagem desses métodos é que, após o agrupamento de dois elementos, eles não são mais separados. Portanto, caso um agrupamento esteja incorreto, ele permanece assim até o final do processamento.

Grande parte dos algoritmos HACMs utilizam uma matriz de similaridades para realizar o agrupamento entre dois elementos. Essa matriz é obtida através do cálculo do valor de similaridade entre todos os elementos do conjunto de dados sobre o qual será realizado o processo de agrupamento.

A seguir são apresentados os quatro principais algoritmos HACMs (ALDENDERFER; BLASHFIELD, 1984).

- **Ligação simples** (*single linkage*): esse método calcula a similaridade entre dois grupos como sendo o maior valor de similaridade entre os elementos pertencentes a esses grupos. Dessa forma, diz-se que esse método utiliza a regra do vizinho mais próximo (MALHOTRA, 2001). Esse método tem uma tendência de encadear elementos criando grupos alongados o que o torna inapropriado para isolar grupos fracamente separados.
- **Ligação completa** (*complete linkage*): diferentemente do método de ligação simples, esse método calcula a similaridade entre dois grupos como sendo o menor valor de similaridade entre os elementos pertencentes a esses grupos. Dessa forma, diz-se que esse método utiliza a regra do vizinho mais distante. Esse método tem a tendência de gerar agrupamentos menores, compactos e hiper-esféricos, contendo elementos altamente similares (ALDENDERFER; BLASHFIELD, 1984).
- **Ligação pelo valor médio ou ligação mediana** (*average linkage*): esse método calcula a similaridade como sendo a média entre o elemento a ser adicionado ao grupo e os demais elementos já presentes no grupo. O novo elemento é adicionado ao grupo cuja similaridade média for maior. Essa similaridade média pode ser calculada de várias formas, sendo que a forma mais utilizada é a média aritmética (ALDENDERFER; BLASHFIELD, 1984).
- **Método de Ward**: esse método é uma variação dos anteriores e procura unir, a cada iteração, o par de grupos cuja união otimiza a variância entre os grupos, juntando os elementos cuja soma dos quadrados entre eles seja mínima ou que o erro desta soma (denominada ESS - *Error Sum of Squares*) seja mínimo (ALDENDERFER; BLASHFIELD, 1984). Tende a produzir grupos hiper-esféricos de tamanhos muito semelhantes.

Conforme já apresentado, grande parte dos algoritmos HACMs são variações de uma única abordagem: iniciam seu processamento considerando que cada elemento corresponde a um grupo e então unem os pares de grupos mais similares até que sobre apenas um grupo. Essa abordagem pode ser expressa através do Algoritmo 2.1.

Algoritmo 2.1: Algoritmo Genérico para HACMs

1: Calcular a matriz de similaridades.
2: Cada elemento do conjunto é seu próprio grupo.
3: **Enquanto** houver mais de um grupo
4: Agrupar o par de grupos mais similar.
5: Atualizar a matriz de similaridades para refletir a similaridade entre o novo
6: grupo e os grupos originais.
7: **Fim-Enquanto**.

2.1.1.1 Fórmula de Lance-Williams

Visto que cada algoritmo HACM difere dos demais apenas pelo modo como identifica o par de grupos mais similar, Lance e Williams (1966) propuseram uma equação para calcular a distância entre dois grupos *Q* e *R*, em que *R* é formado pela junção dos grupos *A* e *B* (equação 2.1). Nessa equação, *p(R,Q)* é a função de distância, enquanto que m_A, m_B e m_Q é o número de elementos nos grupos A, B e Q, respectivamente.

$$p(R,Q) = \alpha_A p(A,Q) + \alpha_B p(B,Q) + \beta p(A,B) + \gamma |p(A,Q) - p(B,Q)| \qquad (2.1)$$

Essa equação pode ser modificada para acomodar qualquer HACM através do ajuste dos valores de α_A, α_B, β, e γ. Portanto, o que diferencia os métodos de ligação simples, ligação completa, de ligação pelo valor médio e de Ward é o ajuste desses parâmetros, conforme apresentado na Tabela 2.1. Nessa tabela,

Tabela 2.1: Valores de parâmetros para a fórmula de Lance e Williams

HACM	α_A	α_B	β	γ
Ligação simples	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação completa	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ligação pelo valor médio	$\frac{m_A}{m_A + m_B}$	$\frac{m_B}{m_A + m_B}$	0	0
Método de Ward	$\frac{m_A + m_Q}{m_A + m_B + m_Q}$	$\frac{m_B + m_Q}{m_A + m_B + m_Q}$	$\frac{-m_Q}{m_A + m_B + m_Q}$	0

Fonte: TAN; STEINBACH; KUMAR, 2006, p. 524

2.1.2 Métodos Hierárquicos Divisivos

Também conhecidos como métodos de classificação, os métodos hierárquicos divisivos possuem uma abordagem "*top-down*", ou seja, todos os elementos são inicialmente agrupamentos em um único grupo e esse vai sendo dividido (ou partido) em grupos menores até que cada grupo contenha apenas um elemento (ALDENDERFER; BLASHFIELD, 1984; EVERITT; LANDAU; LEESE, 2001). Esses métodos tornam-se computacionalmente ineficientes, pois testam todas as possibilidades de divisão de cada grupo existente.

2.1.3 Métodos de Particionamento Iterativo

Os métodos de particionamento iterativo obtêm uma única partição dos dados ao invés de uma estrutura hierárquica de grupos, como o dendograma gerado pelos algoritmos hierárquicos. Esse método gera os grupos fazendo diversas iterações sobre o conjunto de dados.

O algoritmo mais conhecido dessa categoria é o *k*-médias (*k-means*), em que *k* corresponde ao número de grupos que deve ser gerado pelo algoritmo. Neste algoritmo, o usuário indica o número de grupos desejado e o algoritmo cria um conjunto inicial de grupos. A seguir, é computado um valor médio para cada um dos grupos e o algoritmo analisa a similaridade desse valor médio e todos os elementos a serem agrupados. Assim, cada elemento é associado ao grupo em que ele possua o maior valor de similaridade em relação ao valor médio e o valor médio do grupo é novamente computado. Esse processo é repetido até que os valores médios não se alterem (ALDENDERFER; BLASHFIELD, 1984; JAIN; MURTY; FLYNN, 1999).

A maior vantagem dos algoritmos de particionamento iterativo, segundo Aldenderfer (ALDENDERFER; BLASHFIELD, 1984), está no fato deles corrigirem eventuais problemas de agrupamento inadequado (comum nos HACMs) devido as diversas iterações que eles fazem. Porém, por esse motivo, eles são mais demorados do que os hierárquicos aglomerativos e podem não suportar muitos elementos ou tornar o processo inviável. Além disso, outro problema desse método é a definição, *a priori*, do número de grupos que devem ser gerados.

2.1.4 Análise dos Métodos de Agrupamento

A partir da análise dessas três categorias de métodos de agrupamento, optou-se, neste trabalho, pela utilização dos métodos hierárquicos aglomerativos, visto que eles são os mais populares, inclusive em áreas como Banco de Dados (BD) e Recuperação de Informações (RI). Outra vantagem é que esses métodos não necessitam ter como entrada o número de grupos a serem gerados, como ocorre nos métodos de particionamento. Além disso, a maioria dos métodos hierárquicos são determinísticos, ou seja, a ordem dos elementos não altera o agrupamento gerado como ocorre nos métodos de particionamento (MANNING; RAGHAVAN; SCHÜTZE, 2008).

2.2 Validação do Processo de Agrupamento

Visto que a etapa de validação do processo de agrupamento possui um papel fundamental no processo automático de estimativa de valores de R&P, nesta seção, são apresentados alguns pontos importantes do processo de validação, bem como as medidas utilizadas nos experimentos realizados.

Conforme apresentado na Seção 2.1, o processo de agrupamento por similaridade é um processo não-supervisionado, visto que não existem classes pré-definidas nem exemplos que indiquem as características de agrupamento do conjunto de dados (BERRY; LINOFF, 1996). Assim, os diferentes algoritmos de agrupamento são baseados, apenas em suposições para realizar o particionamento dos dados. Portanto, torna-se difícil avaliar se determinado resultado do processo de agrupamento é válido ou não. Isso porque, para avaliar se determinado resultado está correto, deve-se compará-lo com um resultado conhecido. Como neste caso, a intervenção do usuário não é desejada e, muitas vezes, ele não possui muitas informações sobre os dados, torna-se complexo determinar se o resultado está correto ou não.

Ao utilizar algoritmos de agrupamento, um dos problemas a serem resolvidos é decidir qual o número ideal de grupos a serem gerados para cada conjunto de dados, bem como avaliar a qualidade dos grupos gerados. Ambos os problemas têm sido objeto de estudo de vários grupos de pesquisa (DAVE, 1996, GATH; GEVA, 1989, REZAEE et al., 1998, THEODORIDIS; KOUTROUBAS, 1999, XIE; BENI, 1991). No contexto deste trabalho, cada grupo deve conter elementos que representam um único objeto do mundo real.

Segundo Tan (TAN; STEINBACH; KUMAR, 2006), as principais etapas que devem ser realizadas no processo de agrupamento são:

1) Determinar a tendência do agrupamento de um conjunto de dados, ou seja, identificar se estruturas não-randômicas realmente não existem nos dados;
2) Determinar o número correto de grupos;
3) Avaliar quanto do resultado do processo de agrupamento está adequado aos dados, sem utilizar informações externas;
4) Comparar os resultados de um processo de agrupamento com estruturas externas conhecidas, como por exemplo, um conjunto de classes pré-definidas; e
5) Comprar dois conjuntos de grupos para determinar qual deles é o melhor para determinado conjunto de dados.

Percebe-se que as etapas 1, 2 e 3 não utilizam informações externas, ou seja, são etapas não-supervisionadas, enquanto que a etapa 4 requer informações externas. Já a etapa 5 pode ser executada tanto de forma supervisionada, como também, de forma não-supervisionada.

Conforme a literatura da área, os métodos de validação do processo de agrupamento podem ser distribuídos em três categorias principais: baseados em critérios externos, baseados em critérios internos e baseados em critérios relativos (ALDENDERFER; BLASHFIELD, 1984; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001). A seguir, são apresentadas as principais características dessas categorias, bem como as medidas utilizadas nos experimentos para medir a qualidade do processo de agrupamento.

2.2.1 Validação Baseada em Critérios Externos

Nos métodos de validação baseados em critérios, os grupos gerados pelo processo de agrupamento são avaliados em relação a uma estrutura de classes pré-definida (o critério externo) (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001). Essa estrutura, geralmente, reflete alguma intuição sobre a estrutura dos grupos do conjunto de dados e

deve ser criada por um especialista humano. Assim, no agrupamento ideal, cada grupo deve conter apenas os elementos de uma determinada classe.

Nessa categoria encontram-se as seguintes medidas (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001; TAN; STEINBACH; KUMAR, 2006): entropia, pureza, precisão, revocação, medida F (*F-measure*) ou média harmônica, índice de *Jaccard*, estatística de Rand (*Rand statistic*), índice de Folkes e Mallows, estatística de Huberts (Γ) e estatística de Huberts normalizada. A partir dessas medidas, pode-se determinar se o resultado do processo de agrupamento por similaridade aproxima-se mais da estrutura de classes pré-definida, e, portanto, estabelecer se determinado algoritmo é mais eficiente do que outro.

Segundo Aldenderfer et al. (1984), os critérios externos são os melhores para validar uma solução, pois avaliam os grupos em relação a variáveis não utilizadas para gerar o agrupamento, isto é, o critério externo. O maior problema desse tipo de validação está justamente na dificuldade de se definir o critério externo, ou seja, de se criar uma estrutura de validação, visto que esse processo necessita de um especialista humano com conhecimento *a priori* do conjunto de dados.

A seguir, são apresentadas algumas medidas baseadas em critérios externos. Essas medidas foram utilizadas nos experimentos visando encontrar o grau de correlação entre os resultados obtidos com medidas baseadas em critérios externos e medidas baseadas em critérios internos.

2.2.1.1 Precisão

A precisão é uma medida que indica o percentual de elementos corretamente agrupados de acordo com a estrutura de classes do conjunto de dados, ou seja, é a razão entre o número de elementos de uma determinada classe C_j que pertence a um grupo G_i em relação ao número de elementos do grupo G_i. Assim, a precisão de um grupo G_i em relação a uma classe C_j pode ser calculada pela equação 2.2:

$$precisão(G_i, C_j) = \frac{|G_i \cap C_j|}{|G_i|} \tag{2.2}$$

2.2.1.2 Revocação

A revocação é uma medida que indica o percentual de elementos de uma classe que estão contidos em um grupo, ou seja, a razão entre o número de elementos de uma determinada classe C_j que pertencem a um grupo G_i em relação ao número de elementos da classe C_j. Assim, a revocação de um grupo G_i em relação a uma classe C_j pode ser calculada pela equação 2.3:

$$revocação(G_i, C_j) = \frac{|G_i \cap C_j|}{|C_j|} \tag{2.3}$$

2.2.1.3 Medida F ou Média Harmônica

A medida F é uma medida que calcula a média harmônica ponderada dos valores de revocação e precisão. Dessa forma, o valor calculado por essa medida indica se um grupo contém todos os elementos de uma classe e apenas os elementos dessa classe. O

valor da medida F de um grupo G_i em relação a uma classe C_j pode ser calculado pela equação 2.4:

$$medidaF(G_i, C_j) = \frac{2 \times precisão(G_i, C_j) \times revocação(G_i, C_j)}{precisão(G_i, C_j) + revocação(G_i, C_j)} \quad (2.4)$$

Essa medida assume valores no intervalo [0, 1]. O valor 0 indica que nenhum elemento da classe C_j foi agrupado no grupo G_i e o valor 1 indica que todos os elementos e apenas os elementos da classe C_j estão contidos no grupo G_i. Assim, um agrupamento ideal deve retornar um valor igual a 1.

2.2.2 Validação Baseada em Critérios Internos

Os métodos de validação baseados em critérios internos utilizam apenas as informações contidas nos grupos gerados para realizar a validação dos resultados, isto é, não utilizam informações externas, como a estrutura de classes utilizada pelos métodos baseados em critérios externos. Nessa categoria encontram-se as seguintes medidas: coesão e acoplamento (KUNZ; BLACK, 1995), *silhouette coefficient* (TAN; STEINBACH; KUMAR, 2006; ARANGANAYAGI; THANGAVEL, 2007) e *Cophenettic Correlation Coefficient* (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001). A seguir, são apresentadas as medidas para validação baseadas em critérios internos que foram utilizadas neste trabalho.

2.2.2.1 Coesão

Essa medida calcula o grau de similaridade entre os elementos de um mesmo grupo. Quanto maior a similaridade entre os elementos de um mesmo grupo, maior a coesão desse grupo. A coesão de um grupo *G* pode ser calculada pela equação 2.5 (KUNZ; BLACK, 1995):

$$coesão(G) = \frac{\sum_{i>j} sim(g_i, g_j)}{\frac{m(m-1)}{2}} \quad (2.5)$$

Em que $sim(g_i, g_j)$ corresponde ao valor de similaridade entre os elementos *i* e *j* pertencentes ao grupo *G*, *m* é o número de elementos no grupo *G* e *n* é o número de elementos que não estão no grupo *G*.

2.2.2.2 Acoplamento

O acoplamento mede a similaridade média de todos os pares de elementos, sendo que um elemento pertence ao grupo *G* e o outro elemento não pertence ao grupo *G*. O acoplamento pode ser calculado pela equação 2.6 (KUNZ; BLACK, 1995):

$$acoplamento(G) = \frac{\sum_{i,j} sim(g_i, q_j)}{m \times n} \quad (2.6)$$

Em que $sim(g_i, q_j)$ corresponde ao valor de similaridade entre o elemento *i* pertencente ao grupo *G* e o elemento *j* não pertencente ao grupo *G*, *m* é o número de elementos no grupo *G* e *n* é o número de elementos que não pertencem ao grupo *G*.

2.2.2.3 Silhouette Coefficient

A medida *silhouette coefficient* combina as medidas de coesão e acoplamento em uma única medida. O valor do *silhouette coefficient* do *i*-ésimo elemento pode ser calculado pela equação 2.7 (TAN; STEINBACH; KUMAR, 2006; ARANGANAYAGI; THANGAVEL, 2007):

$$silhouetteCoefficient(i) = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.7}$$

Em que a_i é a distância média entre o *i*-ésimo elemento do grupo e os outros elementos do mesmo grupo e b_i é o valor mínimo de distância entre o *i*-ésimo elemento do grupo e qualquer outro grupo que não contém o elemento. Como pode ser observado, a equação acima utiliza funções de distância para calcular o *silhouette coefficient*. Entretanto, uma abordagem semelhante pode ser utilizada para funções de similaridades. O *silhouette coefficient* de um grupo é a média aritmética dos coeficientes calculados para cada elemento pertencente ao grupo.

2.2.3 Validação Baseada em Critérios Relativos

Os métodos de validação baseados em critérios relativos têm como objetivo encontrar o melhor conjunto de grupos que um algoritmo de agrupamento pode definir a partir de certas suposições e parâmetros. Nesses métodos, a avaliação de um agrupamento é feita através da comparação entre esse agrupamento com outros agrupamentos, gerados pelo mesmo algoritmo, porém com diferentes parâmetros de entrada.

Entre os métodos de validação baseados em critérios relativos estão: Índice *Dunn e Dunn Like* (DUNN, 1974), Índice de Davies–Bouldin (DAVIES; BOULDIEN, 1979), Índices RMSSDT e RS (SUBHASH, 1996), Índice SD e Índices S_Dbw (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

No contexto deste trabalho, a validação do processo de agrupamento é utilizada para determinar qual o agrupamento que melhor particiona o conjunto de dados em análise. Visto que, o objeto deste trabalho é eliminar a intervenção do usuário durante o processo de estimativa, optou-se pela utilização de medidas baseadas em critérios internos, isto é, que não necessitam de informações adicionais sobre o conjunto de dados que está sendo agrupado.

2.3 Estimativa de Valores de R&P de Funções de Similaridade

Os mecanismos de avaliação de consultas tradicionais medem apenas o tempo de resposta e o custo do processamento dessas consultas. Isso se deve ao fato de que, pelo critério de igualdade, apenas os elementos relevantes são recuperados (JÓNSSON; FRANKLIN; SRIVASTAVA, 1998). Entretanto, quando se utiliza consultas por abrangência, torna-se necessário medir, também, a qualidade do resultado dessas consultas, visto que o resultado pode variar de acordo com o limiar e a função de similaridade utilizada. O valor usado como limiar implica diretamente na qualidade dos resultados produzidos por essas consultas. A definição de um limiar baixo pode gerar um retorno de muitos elementos incorretos que a função de similaridade determinou como sendo similares ao objeto de consulta, enquanto que um limiar alto pode gerar

uma ausência de elementos relevantes no resultado da consulta. Além disso, um limiar pode ser adequado quando aplicado a uma determinada função de similaridade e inadequado para outra.

Uma das estratégias utilizadas pelos sistemas de RI para avaliar a qualidade do resultado produzido por mecanismos que utilizam funções de similaridade é o uso de medidas de revocação e precisão (R&P). Revocação indica a fração de elementos relevantes que foram recuperados e a precisão indica quantos desses elementos recuperados estão corretos em função da consulta especificada pelo usuário.

A avaliação da qualidade de uma função de similaridade, através de medidas de R&P, requer a execução de várias consultas. Através de um processo iterativo, pode-se gerar um conjunto de *rankings*, gerados com diferentes limiares previamente definidos, para serem avaliados em termos de R&P. Em cada limiar, a média dos valores de R&P resultante do conjunto de *rankings* avaliados é o valor usado para estimar a qualidade da função de similaridade. Como conseqüência, a partir de um indicativo da qualidade desejada, é possível selecionar o limiar adequado a ser utilizado por determinada aplicação ao realizar consultas por abrangência.

Nesta seção, são apresentados dois métodos de estimativa de valores de R&P para diferentes limiares. Inicialmente, é apresentado o método de estimativa clássico, que se torna inviável na prática devido à forte dependência de um especialista humano. Em seguida, é apresentado um método semi-automático, que minimiza essa intervenção do especialista humano, através da utilização de métodos de agrupamento por similaridade.

2.3.1 Método Clássico de Estimativa de Valores de R&P

No método clássico, os valores estimados de R&P para uma coleção de dados podem ser obtidos através dos seguintes passos:

1. Executar consultas, utilizando uma determinada função de similaridade e considerando cada elemento do conjunto de dados como objeto de consulta, obtendo, dessa forma, um *ranking* para cada elemento;
2. Identificar, para cada *ranking* gerado no passo 1, os elementos relevantes de acordo com o objeto de consulta. Essa identificação é realizada por um especialista humano;
3. A partir da identificação dos elementos relevantes, calcular os valores de R&P - utilizando diferentes limiares de similaridade - sobre os *rankings* gerados no passo 1;
4. Calcular o valor médio de revocação e precisão para todas as consultas do passo 1 e para cada valor de limiar.

Claramente, esse processo de estimativa apresenta dificuldades para uso prático com conjuntos de dados reais. Devido ao grande número de *rankings* gerados, a intervenção um especialista humano para indicar os elementos relevantes de cada *ranking* torna-se impraticável.

2.3.2 Método Semi-automático de Estimativa de Valores de R&P

Visando minimizar a intervenção do especialista humano no processo de estimativa de valores de R&P, foi proposto por Stasiu (STASIU; HEUSER; SILVA, 2005; STASIU, 2007) um método semi-automático de estimativa de valores de R&P para vários limiares baseado em amostras de banco de dados. Esse método recebe como

entrada do especialista humano apenas o número de objetos distintos contidos em cada amostra.

O método semi-automático utiliza duas estratégias para diminuir a dependência do especialista humano:

(i) Uso de um processo de amostragem; e

(ii) Utilização de um processo de agrupamento por similaridade.

Os grupos formados pelo processo de agrupamento por similaridade são usados no cálculo automático de R&P, como forma de identificar os elementos relevantes de cada consulta.

O método semi-automático é composto pelas seguintes etapas:

1. **Processo de amostragem:** uma ou mais amostras são geradas a partir de elementos extraídos do conjunto de dados. Um especialista humano analisa a amostra gerada e informa o número de objetos distintos que a amostra contém.
2. **Agrupamento por similaridade:** nesta etapa são gerados grupos de elementos de forma que cada grupo contenha somente as representações de um único objeto do mundo real. O algoritmo de agrupamento utiliza o valor informado pelo especialista humano como critério de quantos grupos devem ser formados.
3. **Consultas por similaridade:** cada elemento da amostra é utilizado como um objeto de consulta sobre os dados da amostra produzindo um *ranking* para cada elemento.
4. **Cálculo de R&P:** nesta etapa são calculados os valores de R&P através da combinação dos resultados produzidos pelas duas etapas anteriores: (i) dos *rankings* gerados pela consulta por similaridade e (ii) dos grupos formados pelo algoritmo de agrupamento por similaridade. O número de elementos em cada grupo corresponde ao número de variações do mesmo objeto que devem ser retornados na consulta.

Esse método foi implementado na ferramenta FERP (Ferramenta para Estimativa de Revocação e Precisão) (BONATO; STASIU; HEUSER, 2005; BONATO, 2005), resultante de um trabalho de conclusão de curso de graduação.

2.3.3 Considerações sobre o Método Semi-automático

O método semi-automático, que está baseado na estimativa de valores de R&P a partir de amostras do conjunto de dados, requer a intervenção de um especialista humano para indicar quantos objetos estão representados em cada amostra. Dessa forma, além da necessidade de intervenção humana, o método deve gerar amostras com poucos elementos para viabilizar que o especialista consiga identificar quantos objetos distintos estão contidos em cada amostra.

Além disso, embora o número de grupos gerados pelo processo de agrupamento seja igual ao número de objetos distintos, isto não significa que cada grupo formado contenha apenas representações de um único objeto do mundo real, visto que nenhum processo de validação é aplicado sobre o resultado. Nesse sentido, o algoritmo de agrupamento pode, também, não conseguir gerar um número de grupos igual ao número

de objetos distintos informados pelo especialista, sendo necessário utilizar um critério de parada alternativo.

No próximo capítulo será apresentado um método automático de estimativa de valores de R&P.

3 MÉTODO AUTOMÁTICO DE ESTIMATIVA DE VALORES DE R&P PARA FUNÇÕES DE SIMILARIDADE

Na Seção 2.3, foi apresentado um método semi-automático de estimativa de valores de R&P de funções de similaridade a partir de amostras de bases de dados. Esse método visa minimizar a intervenção do especialista humano a qual o método clássico é fortemente dependente. Dessa forma, no método semi-automático, o especialista precisa informar apenas a quantidade de objetos distintos que estão representados em cada amostra. Uma limitação dessa abordagem é que o número de elementos de cada amostra é influenciado pela capacidade do especialista em contar os objetos distintos.

Neste capítulo, será proposto um método automático de estimativa de valores de R&P, ou seja, sem a necessidade de intervenção do especialista humano. De acordo com a avaliação experimental apresentada no Capítulo 4, quando os elementos de um conjunto de dados são agrupados de forma que cada grupo contenha apenas representações de um único objeto do mundo real, a medida de validação de grupos *silhouette coefficient* é maximizada. Dessa forma, ao invés de requerer a intervenção humana, o método automático realiza o agrupamento por similaridade dos elementos do conjunto de dados, utilizando diferentes limiares, e seleciona o conjunto de grupos que atinge o maior valor para a medida *silhouette coefficient*. Esse conjunto de grupos é utilizado na etapa do cálculo de valores de R&P.

3.1 Visão Geral do Método Automático Proposto

O método proposto combina duas estratégias para eliminar a intervenção humana durante o processo de estimativa de valores de R&P:

(i) Uso de algoritmos de agrupamento hierárquicos aglomerativos; e

(ii) Uso de uma medida de validação de grupos baseada em critérios internos.

O uso de algoritmos de agrupamento hierárquicos aglomerativos permite que sejam gerados diferentes agrupamentos a partir de diferentes limiares, sem a necessidade de se informar o número de grupos a serem gerados. Isso ocorre pois, nesse tipo de algoritmo de agrupamento, o número de grupos varia de acordo com o limiar utilizado. Assim, durante o processo, são gerados diferentes agrupamentos e a medida de validação dos grupos permite selecionar qual o agrupamento que melhor particiona o conjunto de dados em análise. Dessa forma, torna-se desnecessária a intervenção do especialista humano durante o processo de estimativa.

O processo automático proposto baseia-se na premissa de que o conjunto de grupos selecionado, de acordo com a medida de validação do processo de agrupamento,

particionou corretamente o conjunto de dados em análise, o que significa que cada grupo representa exatamente um e somente um objeto do mundo real. Conforme os experimentos apresentados no Capítulo 4, quando o valor para a medida *silhouette coefficient* é maximizado, os grupos tendem a conter apenas as instâncias de um mesmo objeto real.

3.2 Etapas do Método Automático Proposto

Cada etapa do método automático gera um artefato como resultado (representada por um retângulo com cantos arredondados na Figura 3.1). A partir desse método, podem ser gerados valores de R&P para diferentes limiares, utilizando diferentes funções de similaridade, sem a necessidade de intervenção humana. A seguir, cada etapa do processo será brevemente explicada conforme a seqüência apresentada na Figura 3.1.

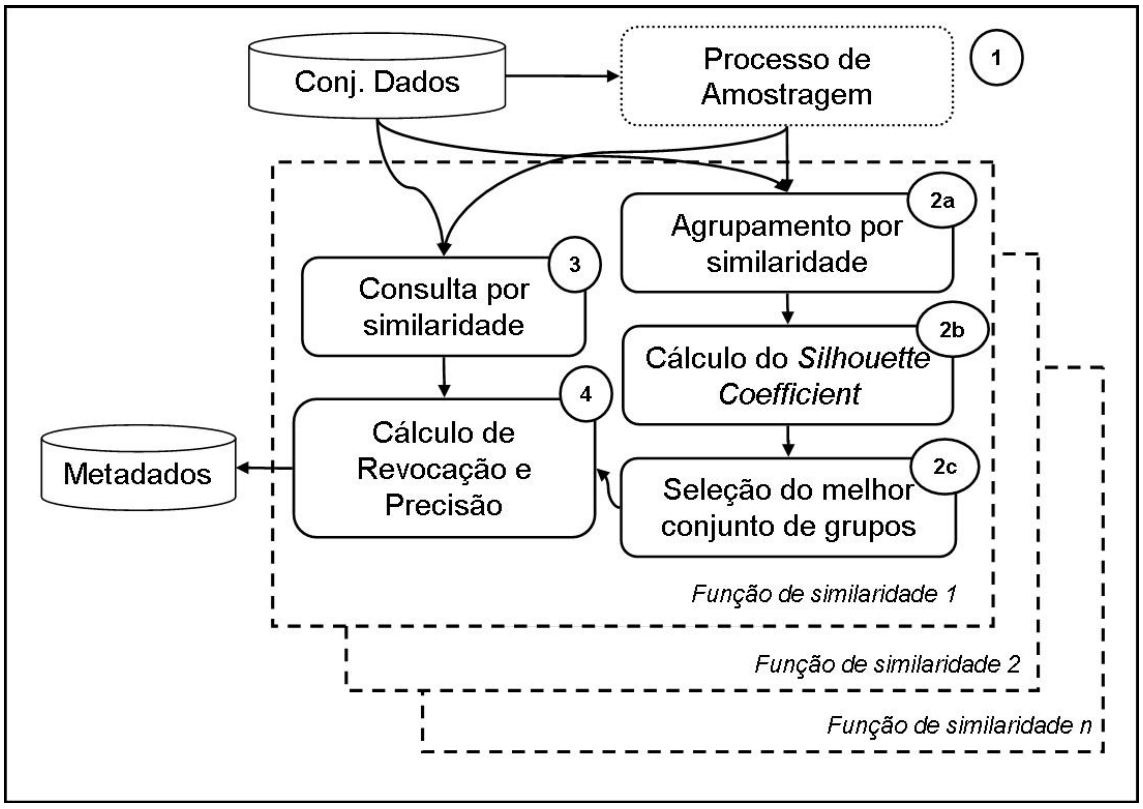


Figura 3.1: Método automático de estimativa de R&P para vários limiares

Processo de amostragem (etapa 1)

Nesta etapa do processo de estimativa, podem ser geradas uma ou mais amostras contendo elementos extraídos do conjunto de dados. Conforme apresentado na literatura sobre estatística (GUERRA; DONAIRE, 1944), pode-se obter uma amostra através de diversos meios. Neste trabalho, foram implementados dois métodos de geração de amostras: aleatório e catação. Ambos os métodos permitem uma distribuição representativa do conjunto de dados.

O processo de amostragem é uma etapa opcional do processo automático de estimativa e visa, apenas, melhorar o desempenho do processo quando estão sendo analisados grandes conjuntos de dados.

Agrupamento por similaridade (etapa 2a)

Esta etapa tem como objetivo gerar grupos de elementos de forma que os elementos de cada grupo sejam representações de um único objeto do mundo real. O critério de agrupamento entre os elementos é determinado por uma função de similaridade. Para facilitar a identificação dos elementos relevantes e irrelevantes para cada consulta, é utilizado um algoritmo de agrupamento que utiliza a mesma função de similaridade que está sendo avaliada.

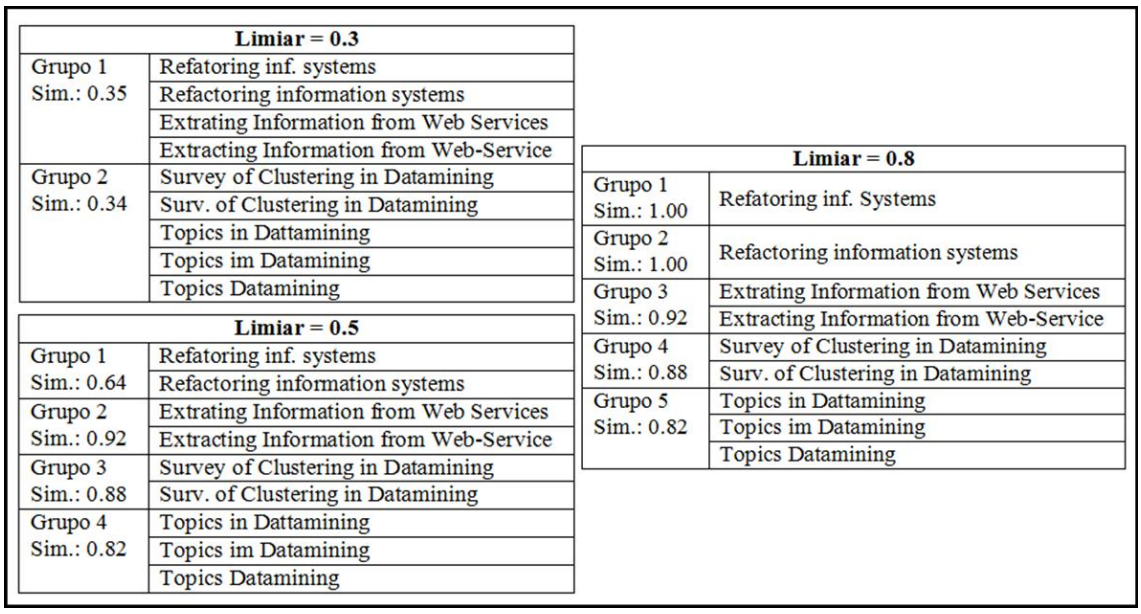
Visto que, os elementos que pertencem a um mesmo grupo correspondem às representações do mesmo objeto real, o número de elementos do grupo ao qual o objeto consultado pertence determina os elementos relevantes que deveriam ser retornados no *ranking* da consulta. Dessa forma, sem que o especialista humano necessite contar e indicar quais são os elementos relevantes, é possível calcular os valores de R&P.

Embora existam diversos tipos de agrupamento, neste trabalho, optou-se por utilizar os algoritmos de agrupamento hierárquicos aglomerativos. Conforme apresentado na Seção 2.1, esses algoritmos não necessitam ter como entrada o número de grupos a serem gerados e são determinísticos, ou seja, a ordem dos elementos não altera o agrupamento gerado (MANNING; RAGHAVAN; SCHÜTZE, 2008). Esses algoritmos produzem como resultado uma árvore binária de elementos, chamada de dendograma, em que cada nodo interno representa um valor de similaridade. Os grupos são formados de acordo com um valor de limiar pré-definido. Portanto, o número de grupos resultante de um processo de agrupamento por similaridade varia de acordo com o limiar utilizado nesse processo. Maiores detalhes sobre a formação dos grupos a partir de um dendograma e utilizando diferentes limiares podem ser obtidos na Seção 2.1

No método semi-automático de estimativa, o processo de agrupamento é executado utilizando um limiar inicial. Caso o número de grupos gerados seja igual ao número de objetos distintos informados pelo especialista humano, o processo de agrupamento é finalizado. Caso contrário, o limiar é ajustado e o processo de geração de grupos é executado novamente. Portanto, esse processo é repetido até que o número de grupos encontrado seja igual ao número de elementos informados pelo usuário. Nesse caso, o processo de agrupamento por similaridade gera um número de grupos igual ao número de objetos distintos representados em cada amostra da coleção. Cabe à função de similaridade, implementada no algoritmo de agrupamento, determinar quais elementos de uma amostra pertencem a um mesmo grupo.

Já no método automático proposto, em que não existe a informação do número de objetos distintos presentes no conjunto de dados, o processo de agrupamento é executado para 11 limiares pré-definidos (0.0, 0.1, 0.2, ..., 1.0), gerando números de grupos diferentes para cada limiar. Na etapa seguinte do processo, os conjuntos de grupos gerados para cada limiar são avaliados por uma medida de validação de grupos e o conjunto de grupos que obter o maior valor é considerado como o conjunto de grupos ideal. A Figura 3.2 apresenta um exemplo de grupos gerados para um conjunto de nomes de títulos de livros utilizando os limiares 0.3, 0.5 e 0.8. Pode-se observar que o

número de grupos varia de acordo com o limiar utilizado. Além disso, cada grupo possui o valor de similaridade (*sim*) do nodo que foi comparado com o limiar.



Limiar = 0.3	
Grupo 1 Sim.: 0.35	Refatoring inf. systems
	Refactoring information systems
	Extrating Information from Web Services
	Extracting Information from Web-Service
Grupo 2 Sim.: 0.34	Survey of Clustering in Datamining
	Surv. of Clustering in Datamining
	Topics in Dattamining
	Topics im Datamining
	Topics Datamining

Limiar = 0.5	
Grupo 1 Sim.: 0.64	Refatoring inf. systems
	Refactoring information systems
Grupo 2 Sim.: 0.92	Extrating Information from Web Services
	Extracting Information from Web-Service
Grupo 3 Sim.: 0.88	Survey of Clustering in Datamining
	Surv. of Clustering in Datamining
Grupo 4 Sim.: 0.82	Topics in Dattamining
	Topics im Datamining
	Topics Datamining

Limiar = 0.8	
Grupo 1 Sim.: 1.00	Refatoring inf. Systems
Grupo 2 Sim.: 1.00	Refactoring information systems
Grupo 3 Sim.: 0.92	Extrating Information from Web Services
	Extracting Information from Web-Service
Grupo 4 Sim.: 0.88	Survey of Clustering in Datamining
	Surv. of Clustering in Datamining
Grupo 5 Sim.: 0.82	Topics in Dattamining
	Topics im Datamining
	Topics Datamining

Figura 3.2: Exemplos de grupos gerados por diferentes limiares

Conforme já mencionado, o método de estimativa de valores de R&P assume que cada grupo contém somente as instâncias de um único objeto real. Entretanto, como a qualidade da função de similaridade é um fator que determina a qualidade do agrupamento, podem existir casos em que somente um especialista humano pode agrupar os elementos corretamente.

Cálculo do *Silhouette Coefficient* (etapa 2b)

O objetivo desta etapa é verificar, através de uma medida de validação baseada em critérios internos, qual o conjunto de grupos que contém apenas elementos que representam um único objeto do mundo real. Como medida de validação, foi escolhida o *silhouette coefficient*, que combina em um único valor, as medidas de coesão e acoplamento. Maiores detalhes sobre o cálculo do *silhouette coefficient* podem ser encontrados na Seção 2.2.

Nessa etapa, para todos os agrupamentos gerados a partir da utilização de diferentes limiares, são calculados os valores do *silhouette coefficient*. Conforme explicado anteriormente, o *silhouette coefficient* utiliza apenas as informações do próprio agrupamento e da matriz de similaridades utilizada pelo algoritmo de agrupamento, sem a necessidade de informações externas.

Conforme apresentado nos experimentos da Seção 4.3, o *silhouette coefficient* apresenta um alto grau de correlação com a medida F, uma medida de validação baseada em critérios externos. Portanto, para cada um dos conjuntos de grupos gerados para uma determinada função de similaridade em análise, é calculado o valor do *silhouette coefficient*. A Tabela 3.1 apresenta um exemplo de valores do *silhouette coefficient* para grupos formados por diferentes limiares.

Tabela 3.1: Valores do *silhouette coefficient* para grupos gerados com diferentes limiares

Agrupamento	***Número de Grupos***	***Limiar***	***Silhouette Coefficient***
1	1	0	0
2	4	0.1	0.287178817
3	12	0.2	0.610821952
4	14	0.3	0.636588346
5	15	0.4	0.591649329
6	17	0.5	0.531745927
7	24	0.6	0.40798929
8	49	0.7	0.218560054
9	94	0.8	0.151040346
10	161	0.9	0.079262754
11	197	1	0

Seleção do melhor conjunto de grupos (etapa 2c)

Nesta etapa, o agrupamento que apresentar o maior valor do *silhouette coefficient* é selecionado como o conjunto de grupos que melhor particiona o conjunto de dados em análise e será utilizado na fase de cálculo de valores de R&P. Por exemplo, de acordo com os valores apresentados na Tabela 3.1, seria selecionado o agrupamento de número 4, formado com o limiar igual a 0.3 visto que, nesse caso, esse agrupamento possui o maior valor para o *silhouette coefficient*.

Consultas por similaridade (etapa 3)

Esta etapa é idêntica à etapa de consulta por similaridade do método semi-automático, isto é, cada elemento do conjunto de dados é utilizado como objeto de consulta sobre o conjunto de dados. Utilizando um mecanismo de consulta que utiliza uma função de similaridade, obtém-se como resultado uma lista de elementos (*ranking*) ordenados de acordo com o seu valor de similaridade (escore) em relação ao objeto consultado. Visto que nesta etapa, todos os elementos do conjunto de dados são utilizados como objeto de consulta, o número de *rankings* gerados é igual ao número de elementos do conjunto de dados.

A Figura 3.3 apresenta exemplos de *rankings* gerados tendo como objeto de consulta os títulos de livros "*XML Database*" e "*Approximate XML Joins*" e utilizando a função de similaridade *Levenshtein* (LEVENSHTEIN, 1966). Ao lado de cada elemento retornado encontra-se o seu escore em relação ao objeto consultado. Como pode ser observado, o *ranking* retornado encontra-se em ordem decrescente do valor do escore. Além disso, o objeto consultado está presente em todos os *rankings*, assim, todas as

consultas têm pelo menos um elemento retornado com escore igual a 1. Em ambos os *rankings* apresentados na Figura 3.3, um limiar de 0.7, por exemplo, seria adequado para separar os elementos relevantes à consulta dos demais.

Ranking 1		
XML Database		
1	XML Database	1.00
2	XML Databases	0.92
3	XML Datebase	0.92
4	XML Databasess	0.86
5	XnL Databases	0.85
6	XML Databasas	0.85
7	XML Daatbases	0.77
8	XML in Databases	0.75
9	XML and Databases	0.71
10	Query XNL in Database	0.52
11	Access Methods for Database	0.37
12	Topics Datamining	0.29
13	Principles of Adm. of Databases	0.29
14	Topics in Dataminig	0.26
15	Topics in Dataminig	0.25
16	SQL and XQL	0.25
17	Topics in Dattamining	0.24
18	Survey of Clustering in Datamine	0.18

Ranking 2		
Approximate XML Joins		
1	Approximate XML Joins	1.00
2	Aproximate XML Joins	0.95
3	Approximate SML Joins	0.95
4	Approximate Join	0.76
5	Approximmate Joins	0.76
6	Similarity Joins	0.38
7	Simmilarity Joins	0.38
8	Architecture for desing paterns implementations	0.23
9	Extrating Information from Web Services	0.23
10	Architecture for design patterns implementations	0.23
11	Exploring structure of data of DICOM files	0.21
12	Exploring strutured data of DICOM	0.21
13	Refactory information systems	0.21
14	Extracting Information from Web-Service	0.21
15	Architeture to design pattern implementation	0.20
16	Refactoring information systems	0.19
17	Suvrey Access Methods	0.19
18	Principles of Administration of Database	0.17

Figura 3.3: Exemplos de *rankings* gerados pela etapa de consulta por similaridade

Cálculo de R&P (etapa 4)

Esta etapa utiliza os seguintes resultados produzidos pelas etapas anteriores:

(i) Os diversos *rankings* gerados pela consulta por similaridade (etapa 3), e

(ii) O agrupamento selecionado pela medida de validação de grupos (etapa 2c).

Para calcular os valores de R&P sem a necessidade de intervenção do usuário, são utilizados os grupos gerados pelo processo de agrupamento. Assim, o número de elementos contidos no grupo ao qual o objeto de consulta faz parte é considerado o número de elementos relevantes**.**

Para realizar a estimativa de valores de R&P para cada *ranking* gerado pelas consultas por similaridade, são aplicados diferentes limiares sobre esses *rankings*. Os valores definidos como limiares devem estar dentro do intervalo de valores retornados pela função de similaridade utilizada no processamento das consultas. Os valores de limiares adotados neste trabalho são os mesmo adotados pelos sistemas de RI. Em sistemas de RI utilizam-se 11 pontos, dentro do intervalo [0.0, 0.1], formando o seguinte conjunto de valores de limiares L = {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. A Figura 3.4 exemplifica os diferentes conjuntos de elementos retornados quando aplicados os limiares 0.3, 0.5 e 0.7 ao *ranking* 1 apresentado na Figura 3.3.

XML Database – Limiar = 0.3		
1	XML Database	1.00
2	XML Databases	0.92
3	XML Datebase	0.92
4	XML Databasess	0.86
5	XnL Databases	0.85
6	XML Databasas	0.85
7	XML Daatbases	0.77
8	XML in Databases	0.75
9	XML and Databases	0.71
10	Query XNL in Database	0.52
11	Access Methods for Database	0.37

XML Database – Limiar = 0.5		
1	XML Database	1.00
2	XML Databases	0.92
3	XML Datebase	0.92
4	XML Databasess	0.86
5	XnL Databases	0.85
6	XML Databasas	0.85
7	XML Daatbases	0.77
8	XML in Databases	0.75
9	XML and Databases	0.71
10	Query XNL in Database	0.52

XML Database – Limiar = 0.7		
1	XML Database	1.00
2	XML Databases	0.92
3	XML Datebase	0.92
4	XML Databasess	0.86
5	XnL Databases	0.85
6	XML Databasas	0.85
7	XML Daatbases	0.77
8	XML in Databases	0.75
9	XML and Databases	0.71

Figura 3.4: Exemplos de *rankings* gerados por diferentes limiares.

Após a aplicação de um limiar sobre um determinado *ranking*, são calculados os valores de R&P. Por exemplo, na Figura 3.4 ao utilizar um valor de limiar igual a 0.3 foram retornados 12 elementos. Entretanto, desses 12 elementos apenas 8 são relevantes a consulta. O valor da precisão, nesse caso, é igual a 0.67. Já o valor de revocação é igual a 1.0, visto que todos os elementos relevantes à consulta foram recuperados. Esse é um exemplo em que foram retornados elementos "falsos positivos", devido a definição de um limiar muito baixo para a função de similaridade em avaliação. Nesse caso, o limiar 0.7 retorna valores de revocação e precisão iguais a 1.

Entretanto, um único *ranking* não pode ser considerado como representativo na avaliação da qualidade de uma função de similaridade. Por esta razão, para cada limiar, são calculados valores médios de revocação e precisão, a partir dos valores de R&P obtidos para cada *ranking* gerado na etapa de consulta por similaridade. Esses valores médios são calculados utilizando a média aritmética.

Como resultado desta etapa, obtém-se uma tabela de valores estimados de R&P para cada um dos limiares. A Figura 3.5 apresenta um exemplo desses valores. Conforme esperado, à medida que o valor do limiar aumenta, o valor de revocação diminui e o valor de precisão aumenta. Os valores de R&P representam a qualidade da função de similaridade sobre determinado domínio de dados. Quanto mais altos os valores tanto para precisão quanto para revocação, melhor a qualidade da função de similaridade.

Limiar	*Revocação*	*Precisão*
0	1.0000	0.1145
0.1	1.0000	0.2566
0.2	0.9989	0.3937
0.3	0.9868	0.5851
0.4	0.8910	0.8701
0.4	0.8257	0.9807
0.6	0.7837	1.0000
0.7	0.6396	1.0000
0.8	0.3967	1.0000
0.9	0.1029	1.0000
1	0.0351	1.0000

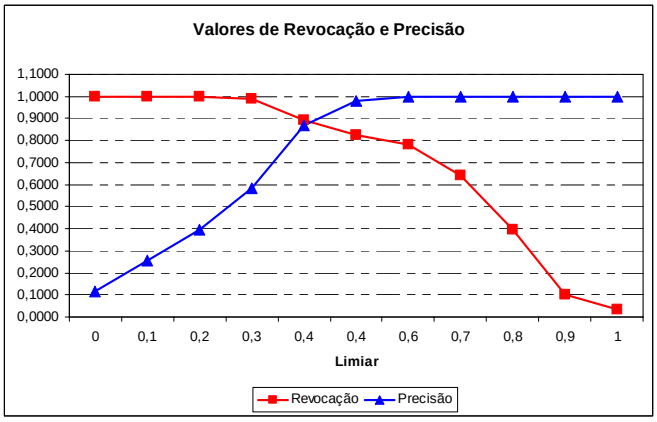


Figura 3.5: Exemplos de representação dos valores de R&P em vários limiares.

4 ANÁLISE EXPERIMENTAL

Neste capítulo, estão descritos os resultados dos experimentos realizados para avaliar o método de estimativa automático proposto no Capítulo 3. Os experimentos realizados estão divididos em três conjuntos. O primeiro conjunto calcula a correlação entre os valores de qualidade do processo de agrupamento obtidos com uma medida baseada em critérios externos (medida F) e uma medida baseada em critérios internos (*silhouette coefficient*). O segundo conjunto avalia os valores de R&P de treinamento estimados pelo método automático a partir de valores de R&P de teste calculados sobre um conjunto de dados de tamanho maior. O último conjunto de experimentos avalia os valores de R&P estimados pelo método automático e os valores de R&P estimados pelo método semi-automático.

Na Seção 4.1 são apresentadas as funções de similaridade usadas nos experimentos. Em seguida, na Seção 4.2 estão descritas as principais características dos conjuntos de dados utilizados nos experimentos. Os procedimentos realizados, bem como os resultados obtidos em cada conjunto de experimentos são apresentados na Seção 4.3.

4.1 Funções de Similaridade Usadas

O grau de semelhança entre um par de elementos pode ser medido através de funções de similaridade ou de distância. Embora o valor numérico obtido por uma função de similaridade possa variar de acordo com a implementação de cada função, nesse trabalho optou-se pela utilização de funções cujo retorno está normalizado no intervalo $[0,1]$. Dessa maneira, considerando que e_1 e e_2 são elementos de um determinado conjunto de dados, uma função de similaridade pode ser definida como $Similaridade(e_1, e_2) \mapsto [0,1]$, sendo que o valor *0* indica que os elementos comparados são totalmente diferentes e o valor *1* indica que são iguais. De acordo com Stasiu (2007), o uso de valores normalizados no intervalo $[0,1]$ facilita a comparação entre os valores retornados por diferentes funções de similaridade.

Com o objetivo de selecionar as funções de similaridade mais adequadas para o domínio de dados utilizado nos experimentos foi utilizada a ferramenta SimEval (HEUSER; KRIESER; ORENGO, 2007). Essa ferramenta realiza uma avaliação da qualidade de diferentes funções de similaridade quando aplicadas sobre um determinado conjunto de dados. Algumas das funções implementadas nessa ferramenta foram

disponibilizadas por Chapman (2007) no projeto *SimMetrics*[1]. Além dessas funções, a ferramenta SimEval também implementa a função de similaridade Carla (MERGEN; HEUSER, 2005) desenvolvida pelo próprio grupo de pesquisa ao qual esse trabalho está vinculado. Para avaliar a qualidade dessas funções, a ferramenta SimEval utiliza a medida da precisão média, conhecida e referência como MAP (*Mean Average Precision*) (SALTON; LESK, 1968), que é uma medida muito utilizada para avaliar a qualidade de sistemas de RI.

Para realizar a avaliação de diferentes funções de similaridade foram executas 30 consultas usando 3 diferentes amostras extraídas de um conjunto de dados que contém registros contidos no mesmo domínio de dados dos conjuntos utilizados nos experimentos. Para cada função de similaridade e amostra utilizada, a ferramenta calculou os seus respectivos valores de MAP. A partir da média aritmética entre os valores de MAP obtidos para cada uma das amostras foram obtidos os valores apresentados na Tabela 4.1.

Tabela 4.1: Avaliação de qualidade de funções de similaridade usando a ferramenta SimEval

Função de Similaridade	***MAP***
Smith-Waterman	0.999512
Q-grams	0.99937
Carla	0.998849
Levenshtein	0.975571
JaroWinkler	0.953562
Jaro	0.951431
Soundex	0.920946
BlockDistance	0.553074
CosineSimilarity	0.553074
DiceSimilarity	0.553074
Jaccard	0.553074

A partir desses resultados obtidos com a ferramenta SimEval foram selecionadas as quatro funções de similaridade que apresentaram o melhor resultado de acordo com a medida de avaliação MAP que são: a *Smith-Waterman* (SMITH; WATERMAN, 1981), a *Q-grams* (NAVARRO, 2001; GRAVANO et al., 2001), a Carla (MERGEN; HEUSER, 2005) e a Levenshtein (LEVENSHTEIN, 1966), Além disso, foi utilizada no primeiro conjunto de experimentos, também, a função de similaridade *Jaccard* (JACCARD, 1912), visto que essa medida apresentou o MAP mais baixo. A função *Jaccard* foi escolhida para realizar uma avaliação das medidas de validação do processo de agrupamento quando o algoritmo de agrupamento utiliza uma função de similaridade

[1] Maiores informações disponíveis em *SimMetrics*: *Similarity Metric Library* (http://sourceforge.net/projects/simmetrics/)

que não separa adequadamente os elementos do conjunto de dados em análise. A seguir são apresentadas algumas características das funções de similaridade usadas nos experimentos:

- **Carla:** essa função retorna um valor que representa o número de caracteres que duas cadeias têm em comum. Essa função é adequada para casos onde duas cadeias possuem muitos termos em comum, sendo que os termos podem aparecer em qualquer lugar dentro das cadeias e podem não possuir espaços entre si. Apesar de essa função apresentar bons resultados na comparação de diversas cadeias, em alguns casos os resultados não são desejáveis. É o caso onde uma inversão na ordem dos termos altera a semântica de uma cadeia. Assim, duas cadeias com significados diferentes seriam consideradas equivalentes.
- ***Levenshtein*:** também conhecida como *Edit Distance*, essa função calcula o número mínimo de operações de inserção, exclusão e substituição de caracteres que são necessárias para fazer com que duas palavras ou cadeias de caracteres fiquem iguais. Nos experimentos realizados, essa função foi implementada para retornar um valor de similaridade ao invés de um valor de distância.
- ***Q-grams*:** essa função calcula a similaridade através da busca por ocorrências dos subconjuntos de uma cadeia de caracteres em outra. Esses subconjuntos correspondem aos *grams* e a quantidade de caracteres em cada *gram* é definida pelo valor de *q*.
- ***Smith-Waterman*:** essa função foi inicialmente desenvolvida para medir a similaridade entre cadeias de proteínas e DNA. Semelhante a função *levenshtein* essa função calcula o custo de transformação de uma seqüência de caracteres em outra.
- ***Jaccard*:** essa é uma função baseada em conjuntos, em que cada elemento desses conjuntos corresponde a uma palavra. Dessa forma a similaridade entre duas cadeias de caracteres *s*1 e *s*2, pela fórmula $|s_1 \cap s_2| \div |s_1 \cup s_2|$.

4.2 Características dos conjuntos de dados

Para realizar os experimentos foram utilizados tanto conjuntos de dados sintéticos quanto um conjunto de dados proveniente de sistemas corporativos reais. Os conjuntos de dados sintéticos foram gerados a partir da ferramenta FEBRL (*Freely Extensible Biomedical Record Linkage*) (CHRISTEN et al. 2004). Essa ferramenta gera conjuntos de dados sintéticos para serem utilizados por sistemas de ligação de registros e de deduplicação. As razões pelas quais foram utilizados conjuntos de dados sintéticos nos experimentos são as seguintes:

(i) Os conjuntos de dados gerados pelo FEBRL já estão previamente classificados. Dessa forma, torna-se possível calcular medidas de avaliação de grupos baseadas em critérios externos. Essa avaliação da qualidade dos grupos baseada em critérios externos é necessária para validar se os valores obtidos com a medida de avaliação baseada em critérios internos são satisfatórios.

(ii) Pode-se controlar alguns parâmetros do conjunto de dados, tais como o número de objetos reais representados (elementos originais) e o número máximo de elementos que representam esses objetos (elementos duplicados).

A tabela 4.2 apresenta um fragmento de um dos conjuntos de dados gerados pelo FEBRL. A primeira coluna contém o identificador de cada item. Esse identificador contém informações que indicam se é o registro corresponde a um elemento original ou a um elemento duplicado. Além disso, esse identificador indica a que classe pertence cada elemento. Por exemplo, o nome "*Kyle ryan"* cujo identificador é "*rec-1-org"* corresponde a um elemento original que está contido na classe *1*, já o nome "*Kyle ryah*" cujo identificado é "*rec-1-dup-1"* corresponde a um elemento duplicado que também faz parte da classe *1*. Dessa forma, cada classe de elementos possui todos os elementos originais e duplicados que representam um único objeto real. Assim, através desses identificadores é possível saber a qual classe pertence cada elemento tornando possível que sejam calculadas medidas baseadas em critérios externos tais como revocação, precisão e medida F.

Tabela 4.2: Exemplos de registros gerados pelo FEBRL

rec_id	**nome**	**rec_id**	**Nome**
rec-1-org	Kyle ryan	rec-2-org	nicholas akot
rec-1-dup-1	Kyle ryah	rec-2-dup-1	nichlas akot
rec-1-dup-2	kile ryan	rec-2-dup-2	nicholas aokt

Para realizar os experimentos de estimativa de valores de R&P foram gerados quatro conjuntos de dados sintéticos de treinamento. O número médio de elementos em cada conjunto varia sistematicamente para permitir uma comparação da qualidade da medida de validação quando aplicada a grupos que contenham poucos elementos, bem como grupos que contenham muitos elementos. O domínio de dados sobre os quais os conjuntos foram gerados é de nomes de pessoas.

Tabela 4.3: Principais características dos conjuntos de dados sintéticos de treinamento

Nome do conjunto de dados	***tNomes1***	***tNomes2***	***tNomes3***	***tNomes4***
Número de objetos reais	100	30	15	7
Número de elementos no conjunto	200	200	200	200
Número médio de elementos por grupo	2	6.67	13.33	28.57

A tabela 4.3. mostra detalhes dos quatro conjuntos de dados gerados pelo FEBRL. A linha "número de objetos reais" indica o número de elementos que realmente se referem a objetos distintos do mundo real, ou seja, são os elementos originais. Dessa forma, o número de grupos gerados para cada conjunto deve ser igual ao número de objetos reais presentes em cada conjunto de dados. A linha "número de elementos no conjunto" refere-se ao número total de elementos contidos em cada conjunto de dados. A

discrepância entre esses dois números corresponde às diferentes representações (elementos duplicados) incluindo os erros gerados pela ferramenta. Já a linha "número médio de elementos por grupo" indica o número médio de elementos que os grupos gerados devem conter para cada conjunto de dados.

Para fins de validação, visando avaliar os valores estimados de R&P com valores de R&P calculados para conjuntos de dados maiores, foram gerados outros quatro conjuntos de dados sintéticos de teste. Esses conjuntos possuem dez vezes o número de registros contidos nos conjuntos utilizados para o cálculo da estimativa. O número médio de grupos, entretanto, permaneceu inalterado. A tabela 4.4 apresenta as principais características desses conjuntos de testes.

Tabela 4.4: Principais características dos conjuntos de dados sintéticos de teste

Nome do conjunto de dados	***vNomes1***	***vNomes2***	***vNomes3***	***vNomes4***
Número de objetos reais	1000	300	150	70
Número de elementos no conjunto	2000	2000	2000	2000
Número médio de elementos por grupo	2	6.67	13.33	28.57

Além dos conjuntos sintéticos gerados pelo FEBRL, também foi utilizado um conjunto de dados de Títulos que pode ser considerado sintético, pois seus registros foram manualmente modificados, sendo criado propositalmente com erros e variações ortográficas. Para a geração desse conjunto de Títulos, foram escolhidos 18 diferentes títulos de artigos científicos de uma base de referências bibliográficas. As modificações (como por exemplo, adição, alteração, remoção e/ou troca de caracteres ou palavras) têm a finalidade de simular possíveis erros de digitação que poderiam ter ocorrido no momento do cadastro de tais títulos. Ao todo, 200 registros de títulos de artigos foram gerados, incluindo os originais.

Além dos conjuntos de dados sintéticos, também foi utilizado nos experimentos um conjunto de dados reais. Esse conjunto contém registros de nomes de cidades. Os registros desse conjunto foram provenientes de um sistema de seleção de candidatos para o vestibular, cujos dados foram fornecidos pelo próprio candidato.

Tanto o conjunto Títulos quanto o de Cidades foram analisados por um especialista humano e para cada elemento foram geradas informações semelhantes àquelas contidas nos registros das coleções geradas pelo FEBRL. Assim, em cada conjunto foram anotados quais os elementos originais e quais os elementos repetidos e a qual classe pertencia cada elemento. Dessa forma, torna-se possível realizar o cálculo de medidas de validação de grupos baseadas em critérios externos também para esses conjuntos de dados. A tabelas 4.5 e 4.6 apresentam exemplos de registros das bases de Títulos e de Cidades, respectivamente. Já a tabela 4.7 apresenta as principais características desses conjuntos de dados.

Tabela 4.5: Exemplos de registros do conjunto de dados de Títulos

rec_id	*nome*
rec-1-org	XML Database
rec-1-dup-0	XML Databases
rec-1-dup-1	XML and Databases
rec-1-dup-2	XML Daatbases

rec_id	*Nome*
rec-2-org	Approximate XML Joins
rec-2-dup-0	Approximate Join
rec-2-dup-1	Approximate SML Joins
rec-2-dup-2	Aproximate XML Joins

Tabela 4.6: Exemplos de registros do conjunto de dados de Cidades

rec_id	*nome*
rec-1-org	CURITIBA
rec-1-dup-0	CUITIBA
rec-1-dup-1	CURITBA
rec-1-dup-2	CURITITBA

rec_id	*Nome*
rec-2-org	PARQUE INDEPENDÊNCIA
rec-2-dup-0	VILA INDEPENDENCIA
rec-2-dup-1	INDEPENDENCIA

Tabela 4.7: Principais detalhes dos conjuntos de dados de Títulos e Cidades

Nome do conjunto de dados	*Títulos*	*Cidades*
Número de objetos reais	18	220
Número de elementos no conjunto	120	300
Número médio de elementos por grupo	6.6	1.36

4.3 Experimentos

Conforme apresentado no Capítulo 3, o método automático de estimativa de R&P tem como principal objetivo eliminar a intervenção do especialista humano do método semi-automático proposto por Stasiu (STASIU; HEUSER; SILVA, 2005; STASIU, 2007). Dessa forma, pode-se observar que as etapas que compõe ambos os métodos são muito semelhantes. A principal diferença do método automático está relacionada com a etapa de agrupamento por similaridade dos elementos. Tendo em vista que, no processo automático não existe a informação de quantos grupos devem ser gerados, a etapa de agrupamento é dividida nas seguintes sub-etapas:

(i) Geração de agrupamentos a partir da execução de algoritmos de agrupamento hierárquicos aglomerativos usando diferentes limiares;

(ii) Avaliação da qualidade dos agrupamentos gerados pelos diferentes limiares através de uma medida de validação baseada em critérios internos; e

(iii) Seleção do agrupamento ideal a partir do valor obtido pela medida de validação do processo de agrupamento.

Além disso, devido a automatização desse processo de estimativa, a etapa de amostragem, que era a etapa inicial do processo semi-automático, tornou-se opcional. No processo automático essa etapa visa apenas melhorar o desempenho do processo de estimativa quando é executado sobre conjuntos com um volume grande de dados, pois o processo validação dos agrupamentos formandos pelos diferentes limiares possui um alto custo computacional.

O método automático proposto assume que o agrupamento, selecionado a partir dos valores obtidos pela medida de validação do processo de agrupamento, dividiu corretamente o conjunto de dados (ou uma amostra desse conjunto). Isso significa que cada grupo contém apenas as representações de um único objeto do mundo real.

Conforme apresentado por Stasiu (2007), os valores estimados de R&P são dependentes do conjunto de dados e da função de similaridade utilizada. Usando funções de similaridades diferentes sobre um mesmo conjunto de dados obtêm-se resultados diferentes. Da mesma forma, usando a mesma função de similaridade sobre conjuntos de dados diferentes produz resultados diferentes. Além disso, o resultado também pode ser dependente do algoritmo de agrupamento utilizado. Portanto, para realizar os experimentos foram utilizados os seguintes algoritmos de agrupamento hierárquicos aglomerativos: algoritmo de ligação simples (SNEATH; SOKAL, 1973), algoritmo de ligação completa (KING, 1967), algoritmo de ligação pelo valor médio (ALDENDERFER; BLASHFIELD; 1984) e algoritmo de Ward (WARD, 1963). Maiores detalhes sobre esses algoritmos podem ser encontrados na Seção 2.1 Nos experimentos realizados, os quatro algoritmos apresentaram desempenhos em termos de tempo e qualidade muito semelhantes. Por isso, nesse trabalho serão apresentados apenas os resultados obtidos com o algoritmo de ligação pelo valor médio que atingiu resultados um pouco melhores. As funções de similaridade utilizadas foram: *Jaccard*, Carla, *Levenshtein*, *Q-grams* e *Smith-Waterman*, já apresentadas na Seção 4.1, sendo que nas figuras serão apresentados apenas os resultados obtidos com as funções de similaridade *Q-grams* e *Smith-Waterman*.

4.3.1 Objetivos

Os experimentos apresentados nessa seção têm os seguintes objetivos:

- Verificar se o agrupamento que melhor particiona o conjunto de dados é aquele que maximiza o valor de uma medida de validação do processo de agrupamento baseada em critérios internos (*silhouette coefficient*);
- Avaliar se os valores de R&P de treinamento estimados através do método automático são semelhantes aos valores de R&P de teste, ou seja, verificar se os valores estimados para um conjunto de dados com poucos elementos podem ser aplicados para conjunto de dados de tamanho maior;
- Avaliar se os valores de R&P estimados pelo método automático são semelhantes aos valores de R&P estimados pelo método semi-automático.

4.3.2 Correlação entre as Medidas F e *Silhouette Coefficient*

Conforme apresentado anteriormente, o primeiro conjunto de experimentos pretende verificar se o melhor agrupamento para um determinado conjunto de dados é aquele que

maximiza o valor de uma medida de validação baseada em critérios internos. Para isso, os experimentos buscam encontrar o grau de correlação entre os valores de validação obtidos pela medida *silhouette coefficient* e os valores de validação obtidos pela medida F.

De acordo com a Seção 2.2, a medida F é uma medida de validação baseada em critérios externos que calcula a média harmônica ponderada de valores de revocação e precisão através da comparação dos grupos gerados com classes pré-definidas. Nos experimentos realizados, para calcular os valores da medida F, foram utilizados os mesmos pesos tanto para revocação quanto para a precisão. Visto que as coleções de dados geradas pelo FEBRL já estão avaliadas, torna-se possível calcular a medida F. Já a medida *silhouette coefficient* é uma medida de validação baseada em critérios internos, isto é, não necessita da comparação dos grupos gerados com classes já definidas. Dessa forma, para o cálculo dessa medida foram utilizadas apenas as informações de similaridade dos elementos contidos nos grupos gerados.

Para esse conjunto de experimentos foram utilizandos os seguintes conjuntos de dados: tNomes1, tNomes2, tNomes3, tNomes4, Títulos e Cidades. Considerando cada função de similaridade avaliada, os passos executados para a realização desses experimentos podem ser resumidos em:

1. Cada conjunto de dados foi agrupado por um algoritmo de agrupamento por similaridade utilizando 11 diferentes limiares (0.0, 0.1, 0.2, ... 1.0). Como resultados, foram gerados 11 diferentes agrupamentos para cada função de similaridade;
2. Para cada agrupamento gerado foi calculado o valor da medida F. Para calcular essa medida foram utilizadas as informações contidas nos registros de cada elemento (apresentados na Seção 4.2). O valor da medida F para um determinado agrupamento é calculado a partir da média aritmética dos valores obtidos para cada grupo que compõe o agrupamento;
3. Para cada agrupamento gerado foi calculado o valor da medida *silhouette coefficient*. Para calcular essa medida foram utilizadas apenas as informações de similaridade entre os elementos dos grupos. Assim como a medida F, o valor da medida *silhouette coefficient* para um determinado agrupamento é calculado a partir da média aritmética do valor dessa medida para cada grupo que compõe o agrupamento.

Visto que foram utilizadas 5 diferentes funções de similaridade e 6 conjuntos de dados, os passos apresentados acima foram executados 30 vezes para cada algoritmo de agrupamento. Como resultado dos passos apresentados, obtém-se uma tabela contendo os valores da medida F e do *silhouette coefficient* para cada agrupamento gerado pelos diferentes limiares. Um exemplo desses valores pode ser visto na tabela 4.8.

Uma etapa importante a ser realizada a partir dos resultados obtidos pelos passos apresentados é calcular o grau de semelhança entre a medida F e o *silhouette coefficient* quanto à qualidade dos agrupamentos gerados. Para calcular essa concordância foi utilizado o coeficiente de correlação de Pearson (LEWICKI; HILL, 2006). O coeficiente de *Pearson* é a medida comum de correlação linear entre dois conjuntos de variáveis contínuas, medindo o grau de associação entre tais variáveis. Esse coeficiente assume valores entre -1 e 1 em que:

- r = 1: significa uma correlação perfeita positiva entre as duas variáveis;

- r = − 1: significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui e;
- r = 0: significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado r = 0 deve ser investigado por outros meios.

O coeficiente de correlação de *Pearson* pode ser calculado pela equação 4.1:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (4.1)$$

Sendo que $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$ são valores medidos de ambas as variáveis. E $\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i-1}^{n} y_i$ são as médias aritméticas de ambas as variáveis.

A tabela 4.8 apresenta alguns valores obtidos a partir da execução dos passos apresentados anteriormente. Esses valores foram obtidos utilizando função de similaridade *Smith-Waterman*. Para os conjuntos de dados tNomes2, tNomes3 e tNomes4 apresentados nessa tabela os limiares que maximizaram os valores para da medida F coincidiram com os limiares que maximizaram os valores do *silhouette coefficient*.

Tabela 4.8: Valores obtidos com as medidas F e *silhouette coefficient*

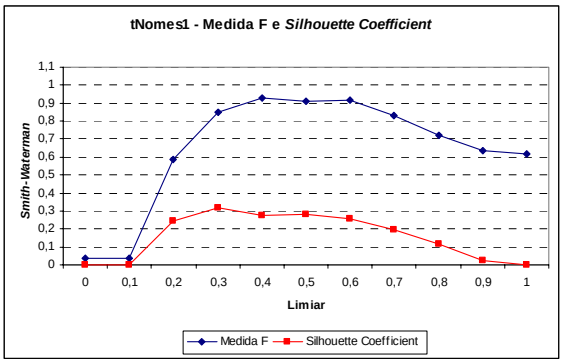
Conj. Dados	*Nro. Grupos*	*Limiar*	*Medida F*	*Silhouette Coefficient*	*Grau de Correlação*
tNomes1	1	0	0.039216	0.000000	0.769018
	1	0.1	0.039216	0.000000	
	22	0.2	0.586074	0.247335	
	52	0.3	0.850264	0.316814	
	73	0.4	0.927644	0.272564	
	88	0.5	0.912170	0.281473	
	101	0.6	0.917162	0.258586	
	126	0.7	0.830234	0.193357	
	159	0.8	0.721953	0.117070	
	191	0.9	0.634505	0.027037	
	200	1	0.614500	0.000000	
tNomes2	1	0	0.104265	0.000000	0.985885
	1	0.1	0.104265	0.000000	
	12	0.2	0.759341	0.415109	
	23	0.3	0.936625	0.467165	
	24	0.4	0.945479	0.478887	
	30	0.5	0.855354	0.374726	

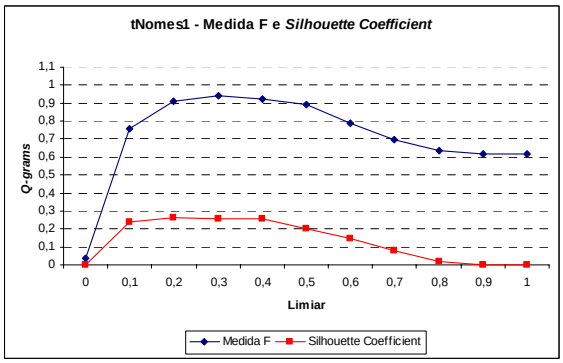
	44	0.6	0.672981	0.290333	
	79	0.7	0.479846	0.187892	
	121	0.8	0.355768	0.148585	
	181	0.9	0.259676	0.051496	
	198	1	0.241420	0.000000	
tNomes3	1	0	0.190045	0.000000	0.970969
	1	0.1	0.190045	0.000000	
	9	0.2	0.854507	0.454381	
	13	0.3	0.982418	0.536734	
	15	0.4	0.960688	0.502456	
	22	0.5	0.708362	0.356209	
	35	0.6	0.510709	0.226152	
	72	0.7	0.303661	0.187367	
	119	0.8	0.209394	0.165869	
	169	0.9	0.156272	0.092073	
	198	1	0.135682	0.000000	
tNomes4	1	0	0.326360	0.000000	0.941585
	1	0.1	0.326360	0.000000	
	4	0.2	0.813195	0.459392	
	7	0.3	1.000000	0.560374	
	7	0.4	1.000000	0.560374	
	13	0.5	0.562392	0.252090	
	20	0.6	0.389698	0.168861	
	45	0.7	0.225755	0.142233	
	96	0.8	0.129008	0.100116	
	165	0.9	0.080512	0.051749	
	198	1	0.068053	0.000000	
Títulos	1	0	0.285714	0.000000	0.984038
	1	0.1	0.285714	0.000000	
	8	0.2	0.845969	0.610443	
	10	0.3	0.892464	0.620641	
	15	0.4	0.921585	0.740206	
	15	0.5	0.921585	0.740206	
	18	0.6	0.909881	0.762634	
	24	0.7	0.800705	0.555192	
	33	0.8	0.624504	0.500724	
	58	0.9	0.414527	0.255298	
	96	1	0.291497	0.000000	
Cidades	1	0.0	0.051948	0.000000	0.463028
	1	0.1	0.051948	0.000000	

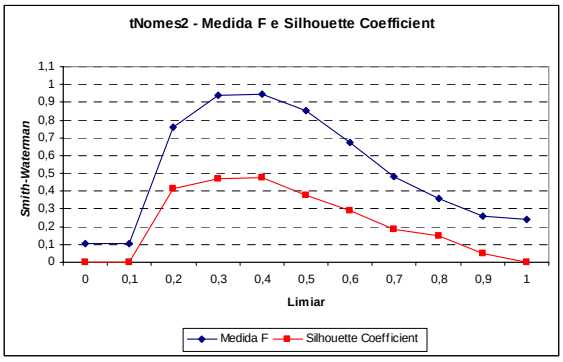
	7	0.2	0.381883	0.183712	
	38	0.3	0.570446	0.224343	
	74	0.4	0.721057	0.220158	
	107	0.5	0.794363	0.229358	
	159	0.6	0.895978	0.204047	
	198	0.7	0.939917	0.160012	
	217	0.8	0.938156	0.132140	
	258	0.9	0.868782	0.065558	

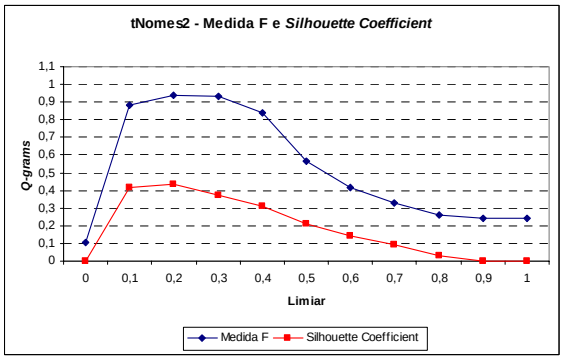
Considerando o critério externo de validação, os valores da medida F geralmente estão entre 0 e 1. Quanto mais altos os valores obtidos com essa medida, maior é a qualidade do agrupamento, pois aumenta a coincidência entre o agrupamento feito pela função de similaridade e o conjunto de classes pré-definidas. Conforme apresentado, todos os conjuntos de dados apresentados na tabela 4.8 obtiveram, para alguns limiares, valores da medida F acima de 0.9. No caso do conjunto de dados tNames4, utilizando limiares iguais a 0.3 e 0.4 o processo de agrupamento conseguiu separar adequadamente os elementos, atingindo um valor igual a 1 para a medida F. Além disso, para esses mesmos limiares foi obtido o maior valor para a medida *silhouette coefficient*.

A Figura 4.1 apresenta 12 gráficos, sendo dois para cada conjunto de dados, utilizando diferentes funções de similaridade, contendo os valores da medida F e do *silhouette coefficient* para 11 limiares. Esses gráficos apresentam os resultados obtidos com as funções de similaridade *Smith-Waterman* e *Q-grams*.









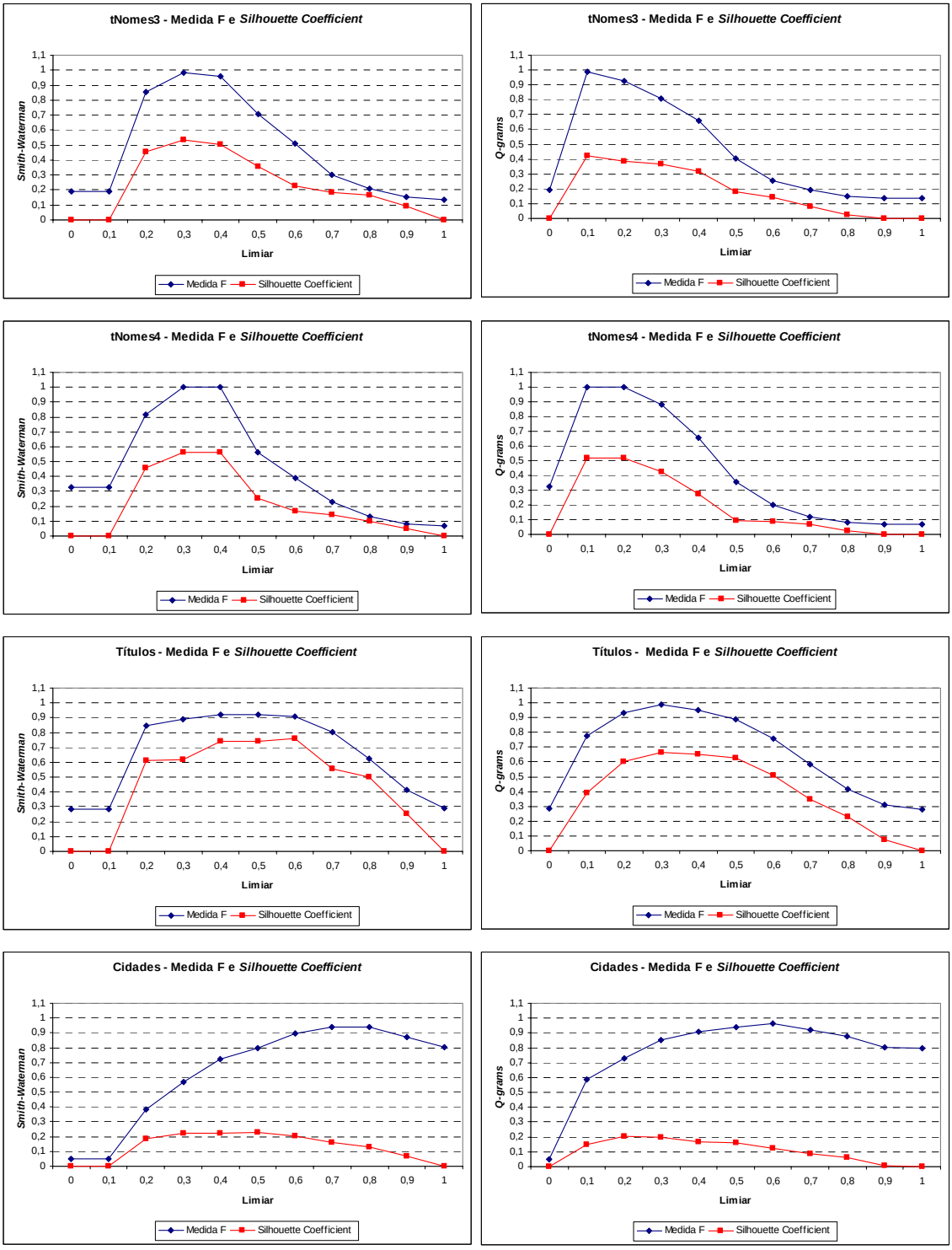


Figura 4.1: Valores das medidas F e *Silhouette Coefficient* para 11 limiares – funções de similaridade *Smith-Waterman* e *Q-grams*

Os gráficos apresentados na Figura 4.1 mostram que em muitos casos o limiar que maximiza a medida F é o mesmo que maximiza o valor do *silhouette coefficient*. Esse fato ocorreu em 12 dos 24 experimentos realizados com as quatro melhores funções de similaridade (6 conjuntos de dados vezes 4 funções de similaridade).

A tabela 4.9 mostra o coeficiente de correlação entre as medidas F e *silhouette coefficient* para todos os conjuntos de dados e as diferentes funções de similaridade utilizadas nos experimentos. Essa tabela mostra que foram obtidos altos graus de correlação entre os conjuntos de dados tNomes2, tNomes3, tNomes4 e Títulos. Já no caso dos conjuntos tNomes1 e Cidades o coeficiente de correlação é menor para todas as funções de similaridade. A razão para isso é que os conjuntos tNomes1 e Cidades contêm grupos relativamente pequenos, com uma média de 2 e 1,36 elementos por grupo, respectivamente. Isso significa que, para esses conjuntos de dados específicos, a função de similaridade deve ser muito precisa e cada valor incorreto gera um impacto relativamente grande na medida de qualidade do agrupamento.

Para os conjuntos de dados tNomes2, tNomes3, tNomes4 e Títulos que obtiveram altos graus de correlação entre as medidas de validação, dos 16 experimentos realizados utilizando as 4 funções de similaridade, em 12 deles o limiar que apresentou o maior valor do *silhouette coefficient* coincidiu com o limiar que apresentou o maior valor da medida F. Considerando os resultados obtidos, pode-se afirmar que é possível utilizar medidas de validação baseadas em critérios internos, no nosso caso o *silhouette coefficient*, para selecionar o agrupamento ideal a ser utilizado no processo de estimativa de R&P, desde que essa medida possua um alto grau de correlação com medidas baseadas em critérios externos. Dessa forma, torna-se possível eliminar a intervenção humana no processo de agrupamento e, consequentemente, do processo de estimativa de valores de R&P.

Tabela 4.9: Grau de correlação entre os valores da medida F e do *silhouette coefficient*

Conj. Dados	***Carla***	***Levenshtein***	***Q-grams***	***Smith-Waterman***
tNomes1	0.604549	0.481762	0.756869	0.769018
tNomes2	0.974983	0.977311	0.985105	0.985885
tNomes3	0.990601	0.983429	0.980967	0.970969
tNomes4	0.941799	0.926048	0.970485	0.941585
Títulos	0.970841	0.962736	0.983037	0.984038
Cidades	0.218789	0.386606	0.463028	0.386575

A Figura 4.2 apresenta 6 gráficos contendo os valores das medidas F e *silhouette coefficient* para 11 limiares utilizando a função de similaridade *Jaccard*. Como pode ser visto nesses gráficos, os valores de validação, baseados no critério externo, obtidos para os grupos gerados por uma função de similaridade que não consegue separar adequadamente os elementos são baixos. Isso significa que os agrupamentos gerados possuem um baixo índice de similaridade com o conjunto de classes pré-definidas. Esse fato ocorre pois, segundo Stasiu (2007), de maneira geral, o processo de agrupamento é dependente da qualidade da função de similaridade. Dessa forma, nos casos em que a funções de similaridade utilizada não é adequada para o domínio de dados em análise, torna-se difícil estimar, de forma automática, os valores de R&P.

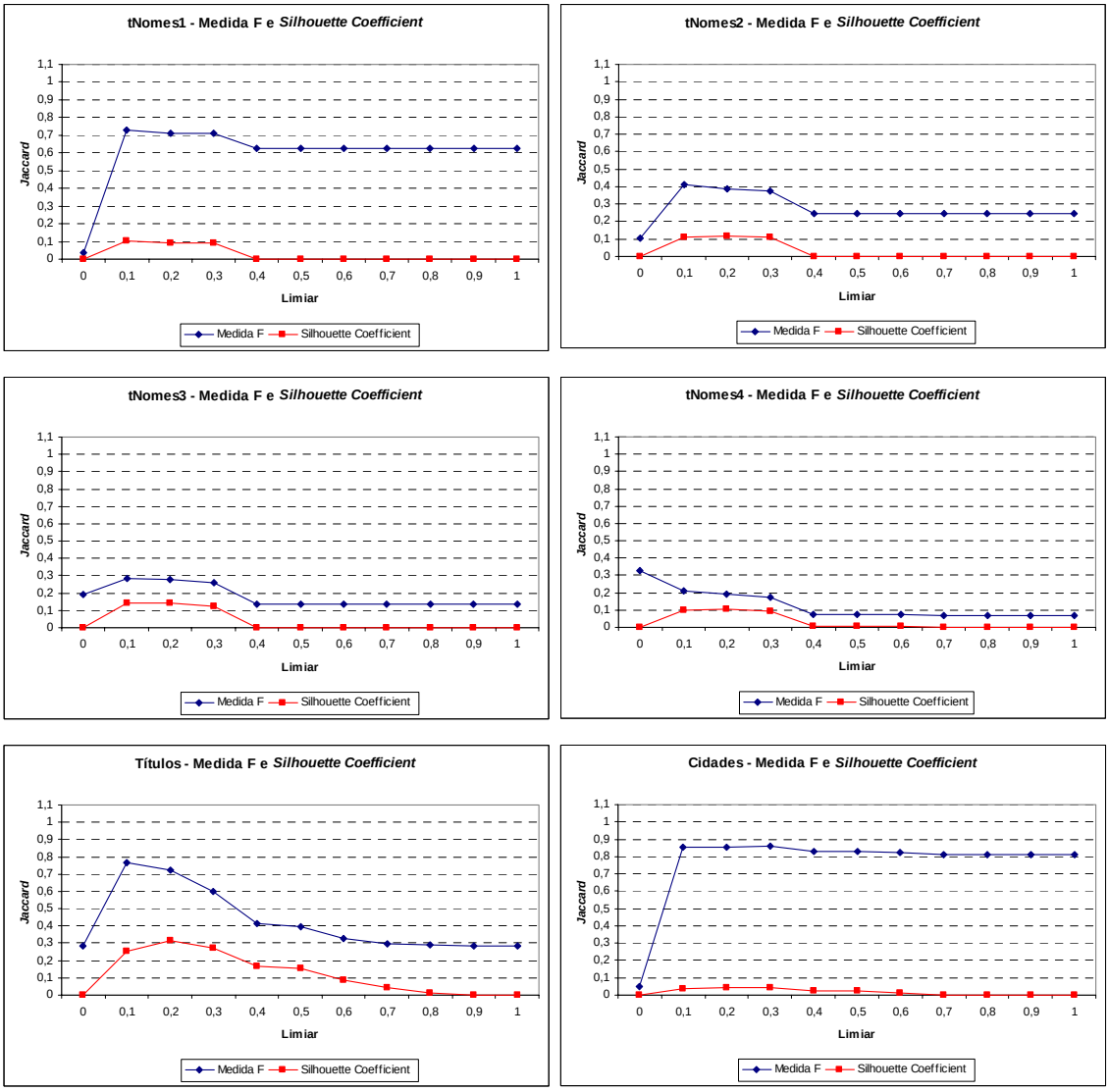


Figura 4.2: Valores das medidas F e *Silhouette Coefficient* para 11 limiares – função de similaridade *Jaccard*

Entretanto, embora os valores obtidos com a medida F sejam baixos quando se utiliza uma função de similaridade inadequada para o domínio dos dados em análise, a correlação entre essa medida e o *silhouette coefficient* continua alta para alguns conjuntos de dados. Os valores de correlação obtidos com os agrupamentos gerados pela função de similaridade *Jaccard* são apresentados na Tabela 4.10.

Tabela 4.10: Grau de correlação entre os valores das medidas F e *silhouette coefficient* utilizando a função de similaridade *Jaccard*

Conj. Dados	***Jaccard***
tNomes1	0.401281
tNomes2	0.870783
tNomes3	0.969356

tNomes4	0.461234
Títulos	0.934129
Cidades	0.582634

4.3.3 Resultado da comparação entre os valores de R&P de treinamento e de teste

O segundo conjunto de experimentos tem como objetivo demonstrar que os valores de R&P de treinamento estimados pelo método automático são muito semelhantes aos valores de R&P de teste calculados utilizado conjuntos de dados maiores.

Considerando cada função de similaridade avaliada os passos executados para a realização do experimento podem ser resumidos em:

1. Cada conjunto de dados de treinamento (*tNomes1*, *tNomes2*, *tNomes3* e *tNomes4*) foi agrupado por um algoritmo de agrupamento por similaridade, utilizando 11 diferentes limiares (0.0, 0.1, 0.2, ... 1.0).
2. Para cada agrupamento gerado foi calculado o valor do *silhouette coefficient*.
3. O agrupamento que apresentou o maior valor para o *silhouette coefficient* foi selecionado como sendo o agrupamento ideal.
4. Cada elemento do conjunto de dados foi usado como um objeto de consulta sobre esse conjunto, gerando um *ranking* para cada elemento.
5. Para cada *ranking* gerado, os valores de R&P foram calculados usando 11 diferentes limiares (0.0, 0.1, 0.2, ... 1.0) utilizando os grupos do agrupamento selecionado para indicar os elementos relevantes e irrelevantes de cada *ranking* aplicado aos diferentes limiares.
6. Para cada limiar, foi calculado a média dos valores de R&P para todas as consultas executadas com esse limiar, obtendo curvas de R&P.
7. De forma semelhante, o processo foi aplicado sobre os conjuntos de teste (*vNomes1*, *vNomes2*, *vNomes3* e *vNomes4*) que contém 10 vezes o número de elementos do conjunto para treinamento. Entretanto, nesse caso, as etapas 2 e 3 foram substituídas pela geração de grupos a partir das informações contidas nos próprios registros do conjunto de dados, isto é, a partir da informação de qual classe corresponde cada elemento. Dessa forma, tem-se certeza que os grupos utilizados para o cálculo dos valores de R&P estão corretos. A partir desses grupos foram calculados os valores de R&P de teste da mesma forma que foram calculados os valores de R&P de treinamento.

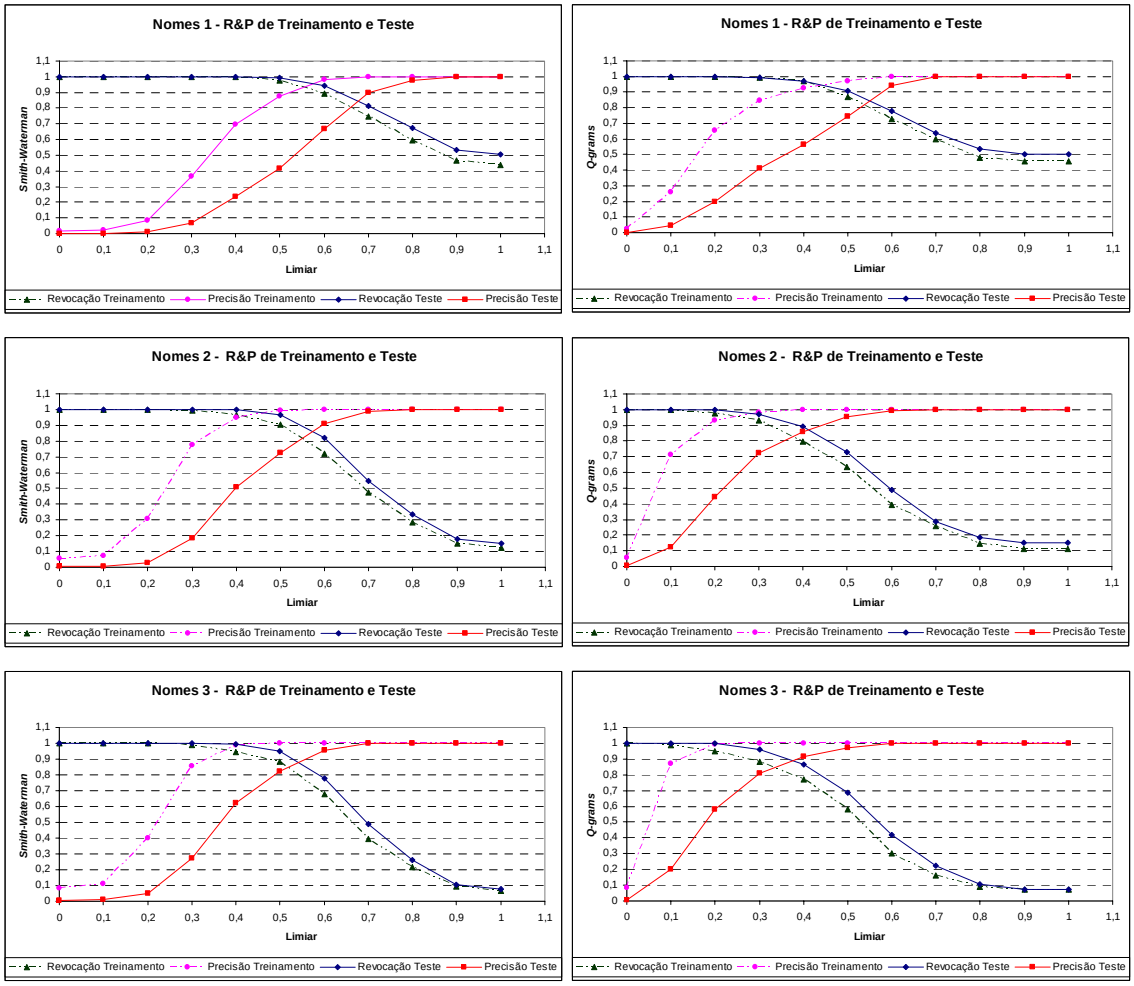
Dessa forma, as curvas de R&P de treinamento são obtidas com o uso de algoritmos de agrupamento por similaridade, portanto, representam os valores estimados. Os valores de R&P de teste foram calculados manualmente utilizando uma estrutura de grupos pré-definida e um conjunto de dados mais representativo, embora com as mesmas características dos conjuntos de dados utilizados para estimar os valores de teste. Por esta razão, é desejável que os valores de revocação e precisão de teste sejam próximos dos valores de revocação e precisão da coleção.

A distância entre os valores de R&P de treinamento e os valores de R&P de teste foi medida através do Desvio Quadrático Médio (MSD - *Mean Square Deviation*), definida pela equação 4.2:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2 \qquad (4.2)$$

Sendo n o número do limiar analisado, $\hat{x}$ corresponde aos valores de R&P de treinamento e x corresponde aos valores de R&P de teste. O MSD tem como objetivo medir a distância entre os pontos de duas curvas. Quanto menor for o valor do MSD significa que os valores estão mais próximos e são mais confiáveis.

Os resultados obtidos para as funções de similaridade *Q-grams* e *Smith-Waterman* são apresentados na Figura 4.3, através das curvas de R&P de treinamento e de teste. A função de similaridade representada em cada gráfico é indicada próximo ao eixo *y*. Esse eixo indica os valores médios, de revocação e precisão, obtidos para cada um dos 11 limiares pré-definidos que correspondem ao eixo *x*.



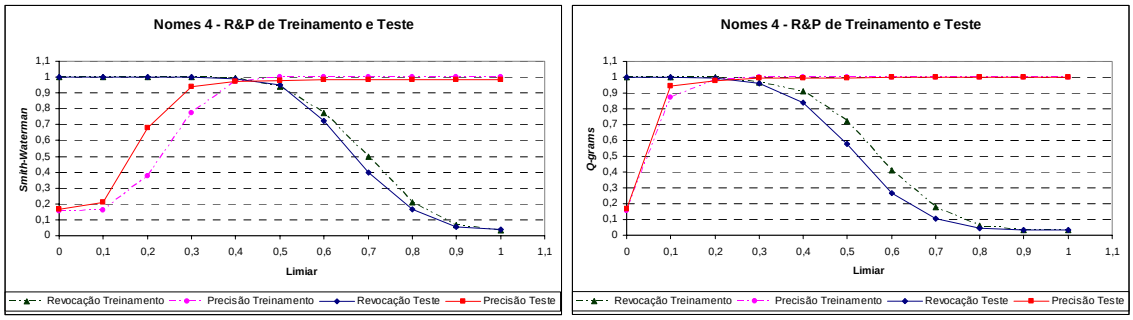


Figura 4.3: Valores de R&P de treinamento e de teste

De acordo com os gráficos apresentados, os valores de R&P gerados pelo método automático de estimativa possuem uma pequena diferença quando comparados com os valores de R&P de teste. As tabelas 4.10 e 4.11 apresentam os valores MSD obtidos pela comparação dos valores de revocação e precisão de teste com os valores de treinamento, respectivamente. Quanto mais altos os valores de MSD, maior é a diferença entre os valores de treinamento e de teste. Como é possível observar, os valores são baixos indicando, que os valores de treinamento estimados são próximos dos valores reais obtidos com os conjuntos de teste. Esses valores são semelhantes aos valores apresentados em Stasiu (2007) em que é feita uma comparação dos valores de estimativa para as amostras geradas e os valores gerados para o conjunto de dados. Esse fato é um indicativo de que a eliminação da intervenção de um especialista humano no processo de estimativa não possuem um impacto significativo no processo de estimativa. Esse fato indica a viabilidade do processo automático de estimativa. Além disso, essas tabelas mostram que o processo automático de estimativa é mais preciso na estimativa de valores de revocação (valores de MSD abaixo de 1.27%) e menos preciso no processo de estimativa de valores de precisão (valores de MSD abaixo de 10.11%).

Tabela 4.11: MSD entre os valores estimados de revocação e os valores reais

	Carla	*Levenshtein*	*Q-grams*	*Smith-Waterman*
Nomes1	0.0074	0.0109	0.0011	0.0020
Nomes2	0.0127	0.0010	0.0031	0.0023
Nomes3	0.0036	0.0051	0.0043	0.0025
Nomes4	0.0086	0.0085	0.0047	0.0013

Tabela 4.12: MSD entre os valores estimados de precisão e os valores reais

	Carla	*Levenshtein*	*Q-grams*	*Smith-Waterman*
Nomes1	0.0562	0.0763	0.0572	0.0574
Nomes2	0.0936	0.0521	0.0615	0.0648
Nomes3	0.1011	0.0658	0.0608	0.0593
Nomes4	0.0362	0.0063	0.0005	0.0113

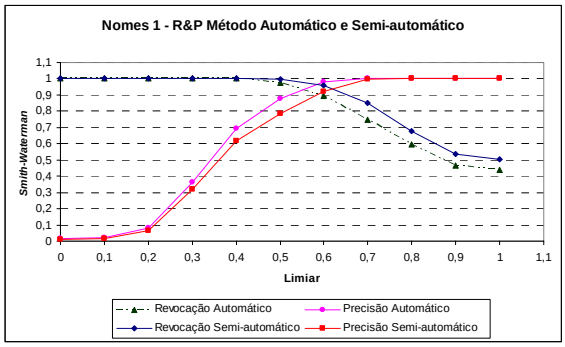
4.3.4 Resultado da comparação dos valores de R&P obtidos pelo método automático e pelo método semi-automático

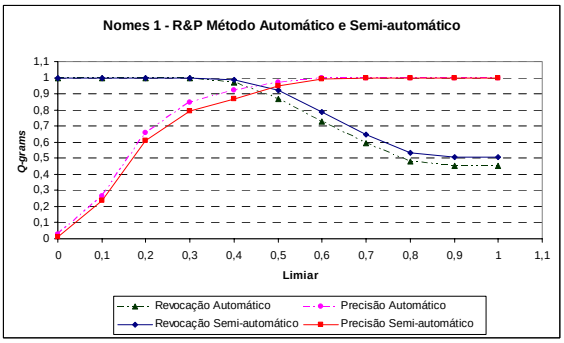
O terceiro conjunto de experimentos tem como objetivo demonstrar que os valores de R&P estimados pelo método automático são muito semelhantes aos valores de R&P estimados pelo método semi-automático. Para esse conjunto de experimentos foram utilizados os valores de R&P obtidos com a execução dos 6 primeiros passos apresentados na subseção 4.3.3. Esses 6 primeiros passos também foram executados para os conjuntos de dados de *Títulos* e Cidades. Após, foram realizadas os passos correspondentes ao processo de estimativa semi-automático que podem ser resumidos em:

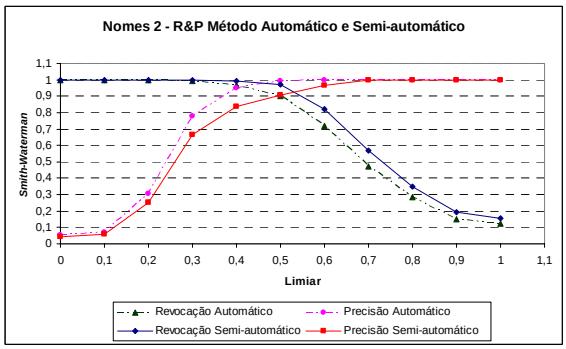
1. Para cada conjunto de dados (*tNomes1*, *tNomes2*, *tNomes3*, *tNomes4*, *Títulos* e *Cidades*), um especialista humano informou o número de objetos distintos contidos em cada um dos conjuntos;
2. Os elementos dos conjuntos foram agrupamentos usando os mesmos algoritmos de similaridade já mencionados e utilizando o número de objetos distintos como critério de parada conforme descrito na Seção 2.3;
3. Cada elemento foi usado como um objeto de consulta sobre o conjunto de dados, gerando um *ranking* para cada elemento.
4. Para cada *ranking* gerado, os valores de R&P foram calculados para 11 diferentes limiares (0.0, 0.1, 0.2, ... 1.0) utilizando o resultado do processo de agrupamento da etapa 2.
5. Para cada limiar, foi computada a média dos valores de revocação e de pecisão para todas as consultas executadas utilizando esse limiar;
6. Os valores de estimativa de R&P gerados pelo método automático foram comparados com os valores de R&P obtidos pelo método semi-automático.

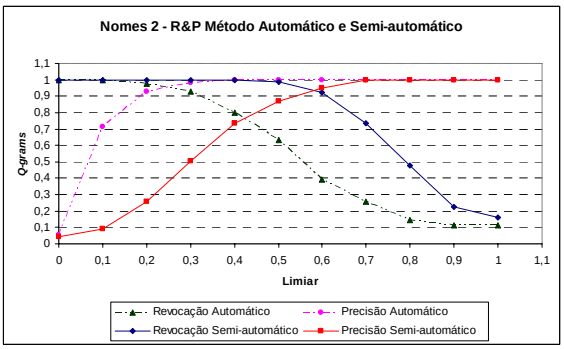
A distância entre os valores de R&P de estimados pelo método automático e os valores de R&P de estimados pelo método semi-automático foi medida através do MSD conforme já definido pela equação 4.2.

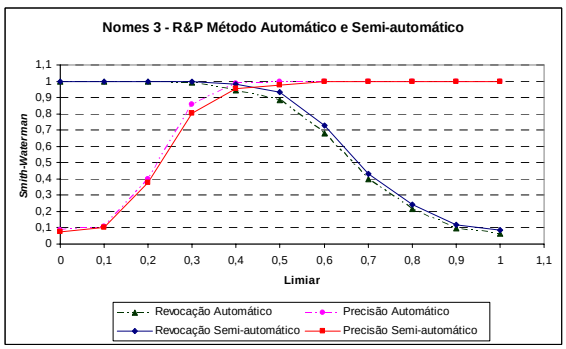
Os resultados obtidos para as funções de similaridade *Q-grams* e *Smith-Waterman* são apresentados na Figura 4.4. Da mesma forma que os gráficos apresentados na Figura 4.3, a função de similaridade representada em cada gráfico é indicada próximo ao eixo *y*. Esse eixo indica os valores médios, de revocação e precisão, obtidos para cada um dos 11 limiares pré-definidos que correspondem ao eixo *x*.

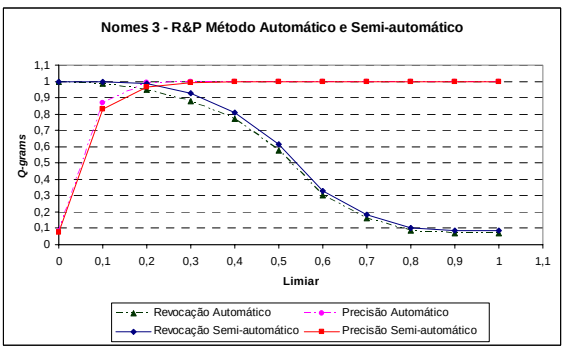

Nomes 1 - R&P Método Automático e Semi-automático
Smith-Waterman
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático

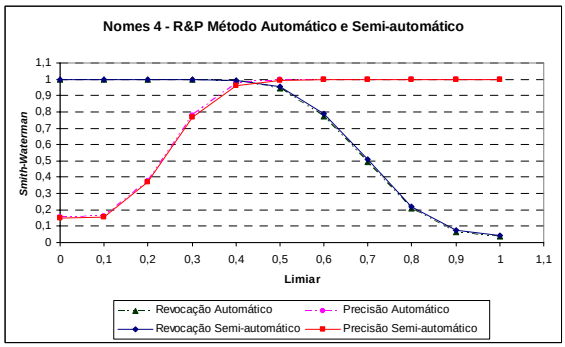

Nomes 1 - R&P Método Automático e Semi-automático
Q-grams
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático

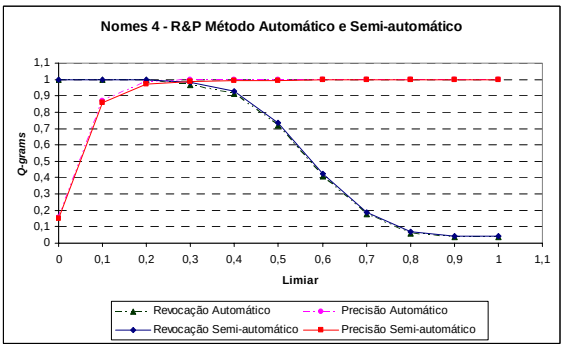

Nomes 2 - R&P Método Automático e Semi-automático
Smith-Waterman
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático

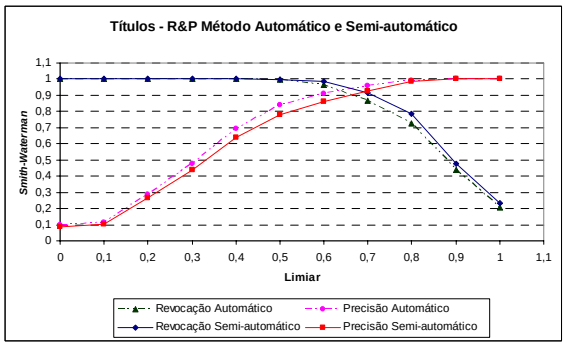

Nomes 2 - R&P Método Automático e Semi-automático
Q-grams
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático

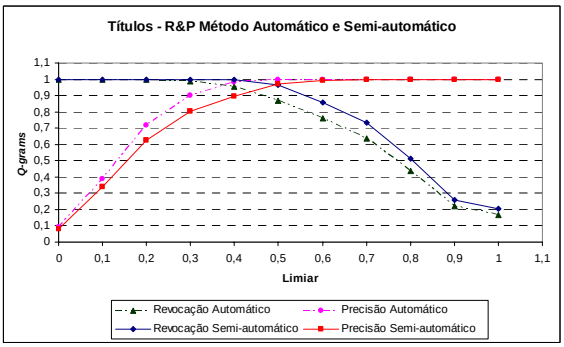

Nomes 3 - R&P Método Automático e Semi-automático
Smith-Waterman
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático


Nomes 3 - R&P Método Automático e Semi-automático
Q-grams
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático


Nomes 4 - R&P Método Automático e Semi-automático
Smith-Waterman
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático


Nomes 4 - R&P Método Automático e Semi-automático
Q-grams
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático


Títulos - R&P Método Automático e Semi-automático
Smith-Waterman
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático


Títulos - R&P Método Automático e Semi-automático
Q-grams
Limiar
Revocação Automático
Precisão Automático
Revocação Semi-automático
Precisão Semi-automático

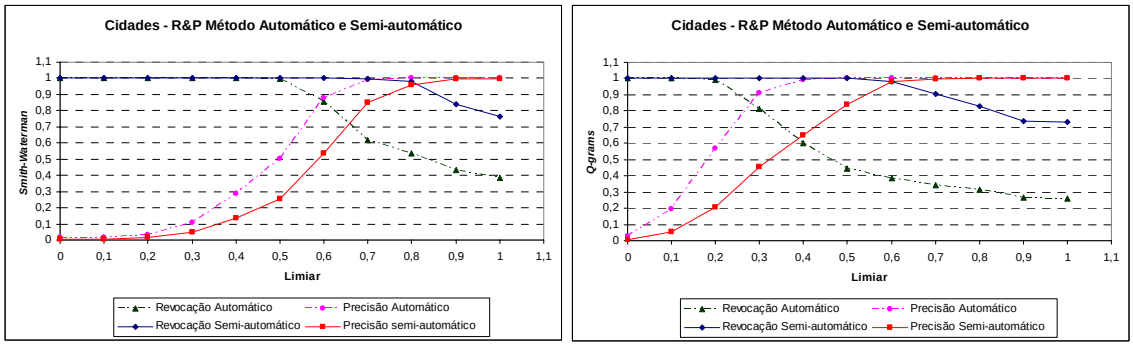


Figura 4.4: Comparação entre os valores de R&P estimados pelo método automático e pelo método semi-automático

De acordo com os gráficos apresentados, os valores de R&P gerados pelo método automático de estimativa possuem uma pequena diferença quando comparados com os valores de R&P gerados pelo método semi-automático. Nos gráficos referentes aos conjuntos de dados Nomes1, Nomes3, Nomes4 e Títulos pode-se observar que as curvas estão praticamente sobrepostas, o que significa que os valores obtidos em ambos os métodos são praticamente iguais.

Essa afirmação pode ser confirmada pelo cálculo do MSD entre essas curvas. As tabelas 4.13 e 4.14 apresentam os valores MSD obtidos pela comparação dos valores, de revocação e precisão, obtidos por ambos os métodos de estimativa. Como pode ser observado, nessas tabelas os conjuntos de dados mencionados anteriormente são aqueles que apresentam os menores valores de MSD tanto para as curvas de revocação quanto para precisão. De acordo com os valores de MSD apresentados por essas tabelas, pode–se afirmar que a eliminação da intervenção de um especialista humano no processo de estimativa não possui um impacto significativo no processo de estimativa de valores de R&P.

Tabela 4.13: MSD entre os valores de revocação estimados pelo método automático e pelo método semi-automático

	Carla	***Levenshtein***	***Q-grams***	***Smith-Waterman***
Nomes1	0.0085	0.0147	0.0015	0.0029
Nomes2	0.0114	0.0017	0.0735	0.0028
Nomes3	0.0005	0.0032	0.0007	0.0008
Nomes4	0.0001	0.0093	0.0001	0.0001
Títulos	0.0003	0.0020	0.0035	0.0008
Cidades	0.1218	0.0966	0.1718	0.0612

Tabela 4.14: MSD entre os valores de precisão estimados pelo método automático e pelo método semi-automático

	Carla	***Levenshtein***	***Q-grams***	***Smith-Waterman***
Nomes1	0.0037	0.0058	0.0008	0.0018
Nomes2	0.0082	0.0018	0.1051	0.0034
Nomes3	0.0007	0.0010	0.0002	0.0004
Nomes4	0.0001	0.0029	0.0001	0.0000
Títulos	0.0002	0.0011	0.0027	0.0011
Cidades	0.0362	0.0311	0.2453	0.0208

5 CONCLUSÃO

A definição do limiar adequado a ser utilizado em consultas por abrangência normalmente está relacionada com a estimativa dos efeitos do uso diferentes limiares na qualidade dos resultados dessas consultas. As consultas por abrangência utilizam funções de similaridade juntamente com um limiar para restringir o resultado, retornando apenas os elementos relevantes, isto é, apenas os elementos que representam o objeto de consulta. O processo clássico de estimativa da qualidade de funções de similaridade baseia-se fortemente na intervenção humana em rotular cada elemento como relevante ou não relevante para determinada consulta.

Nesta dissertação foi proposto um método automático de estimativa de valores de R&P de funções de similaridade para vários limiares. A partir desse método, torna-se desnecessária a intervenção humana durante o processo de estimativa. A abordagem utilizada para automatizar esse processo foi utilizar algoritmos de agrupamento hierárquicos aglomerativos realizados utilizando diferentes limiares combinados com uma medida de validação desse processo de agrupamento. Dessa forma, o agrupamento ideal a ser utilizado na estimativa de valores de R&P era o agrupamento que maximizava essa medida de validação. Como foi utilizada uma medida de validação baseada em critérios internos, isto é, que não necessita ter os dados previamente avaliados, tornou-se possível a eliminação do especialista humano.

Os experimentos realizados demonstram que: (i) existe um alto grau de correlação entre a medida baseada em dados externos (medida f) e a medida baseada em critérios internos (*silhouette coefficient*), dessa forma pode-se utilizar essa medida no processo de validação; (ii) os valores estimados de R&P de treinamento são muito semelhantes aos valores de R&P de teste, visto que as distâncias estatísticas entre esses valores, calculadas pelo método desvio quadrático médio (MSD), são próximas de zero; e (iii) os valores de R&P obtidos pela execução do método automático são muito semelhantes aos valores de R&P obtidos com o método semi-automático, visto que a distância entre esses valores, calculadas, também, pelo MSD, é próxima de zero. Por esses resultados, entre outros realizados durante o desenvolvimento desta dissertação, conclui-se que é viável utilizar o método automático, ou seja, é possível eliminar a intervenção humana durante o processo de estimativa de valores de R&P.

Os experimentos demonstraram, também, que para alguns conjuntos de dados o método produz resultados menos precisos. Isso ocorre, geralmente, com conjuntos de dados que geram grupos com poucos elementos bem como quando se utiliza funções de similaridade que não são adequadas para o domínio dos dados.

5.1 Trabalhos Futuros

Como trabalhos futuros podem ser citadas as seguintes propostas:

- Encontrar uma função, baseada nos valores de validação gerados por critérios internos, que indique qual o melhor agrupamento considerando diferentes funções de similaridade.
- Realizar experimentos utilizando outros algoritmos de agrupamento por similaridade.
- Realizar experimentos utilizando outros conjuntos de dados.

REFERÊNCIAS

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis.** Newbury Park: Sage, 1984. (Sage University Paper Series on Quantitative Applications in the Social Science, 44).

ARANGANAYAGI, S.; THANGAVEL, K. Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND MULTIMEDIA APPLICATIONS, ICCIMA, 2007, Sivakasi, India. **Proceedings...** Los Alamitos: IEEE, 2007. p. 13-17.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval.** New York: Addison Wesley, 1999.

BENJELLOUN, O. et al. Generic entity resolution in the serf project. **IEEE Data Engineering Bulletin**, [S.l.], June 2006.

BERRY, M. J. A.; LINOFF, G. **Data Mining:** Techniques For marketing, Sales and Customer Support. New York: John Willey & Sons, 1996.

BONATO, J.; STASIU, R.; HEUSER, C. FERP: ferramenta para estimativa de revocação e precisão. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 20., SESSÃO DE DEMOS, 2., 2005, Uberlândia, MG, Brazil. **Anais...** [S.l.:s.n.], 2005.

BONATO, J. **Ferramenta para Estimativa de Revocação e Precisão usando Amostras de Banco de Dados.** 2005. 76 f. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

COHEN, W. W. Data integration using similarity joins and a word-based information representation language. **ACM Trans. Inf. Syst.**, New York, NY, USA, v.18, n.3, p.288-321, 2000.

CHRISTEN, P. et al. FEBRL - a parallel open source data linkage system. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PAKDD, 8., 2004, Sydney, Australia. **Advances in Knowledge Discovery and Data Mining**: proceedings. Berlin: Springer, 2004. p. 638-647. (Lecture Notes in Computer Science, v.3056).

CUTTING, D. et al. Scatter/Gather: a cluster-based approach to browsing large document collections. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, SIGIR, 1992. **Proceedings...** New York: ACM Press, 1992. p.318-329.

DAVE, R. N. Validating fuzzy partitions obtained through c-shells clustering. **Pattern Recognition Letters**, Amsterdam, v. 10, p. 613-623, 1996.

DAVIES, D. L.; BOULDIN, D. W. Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 1, n. 2, p. 95-104, 1979.

DUNN, J. C. Well Separated Clusters and Optimal Fuzzy Partitions. **Journal of Cybernetica**, [S.l.], v. 4, p. 95-104, 1974.

EVERITT, B. S. et al. **Cluster Analysis**. 4th ed. New York: Oxford University Press, 2001. 237 p.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. **Journal of the American Statistical Society**, [S.l.], v.64, p.1183-1210, 1969.

GATH, I.; GEVA, A. B. Unsupervised optimal fuzzy clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Amsterdam, v. 11, n. 7, p. 773-780, 1989.

GUERRA, M. J.; DONAIRE, D. **Estatística Indutiva**: teoria e exercícios. São Paulo: LTCE, 1944.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering Algorithms and Validity Measures. In: INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, 13., 2001, Washington, DC, USA. **Proceedings...** Los Alamitos, CA: IEEE Computer Society, 2001. p.3-22.

HARTIGAN, J. A. **Clustering Algorithms.** New York, NY, USA: John Wiley and Sons, 1975.

HEUSER, C. A.; KRIESER, F. N. A.; ORENGO, V. M. SimEval: a tool for evaluating the quality of similarity functions. In: CONFERENCE ON CONCEPTUAL MODELLING, 26., 2007, Auckland, Nova Zelandia. **Chanlenges in Conceptual Modelling**: tutorials, posters, panel and industrial contributions. Sidney: Australian Computer Society, 2007.

JACCARD, P. The Distribution of the Flora in the Alpine Zone. **New Phytologist**, [S.l.], v.11, n.2, p.37-50, Feb. 1912.

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

JÓNSSON, B. T. ; FRANKLIN, M.; SRIVASTAVA, D. Interaction of Query Evaluation and Buffer Management Retrieval. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT DATA, SIGMOD, 1998, Seattle, Washington, USA. **Proceedings...** New York : ACM, 1998. p. 118-129.

KAUFMAN, L.; ROUSSEEUW, P. **Finding Groups in Data**: An Introduction to Cluster Analysis. New York: Wiley, 1990.

KING, B. Step-wise Clustering Procedures. **J. Am. Stat. Assoc.**, [S.l.], v. 69, p. 86-101, 1967.

KUNZ, T; BLACK, J. Using Automated Process Clustering for Design Recovery and Distributed Debugging. **IEEE Trans. Software Engineering**, Los Alamitos, CA, v. 21, n. 6, p. 515-527, 1995.

LANCE, G. N.; WILLIAMS, W. T. A General Theory of Classificatory Sorting Strategies. **Computer Journal**, [S.l.], n. 9, p.373-380, 1966.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**, Woodburry, p. 707–710, 1966.

MALHOTRA, N. K. **Pesquisa de Marketing**: uma orientação aplicada. 3.ed. Porto Alegre: Bookman, 2001. 720 p.

MANNING, C. D.; RAGHAVAN, P; SCHÜTZE, H. **Introduction to Information Retrieval**. [S.l.]:Cambridge University Press, 2008.

MERGEN, S.; HEUSER, C. A. Carla: uma técnica para comparação de cadeias de caracteres. In: ESCOLA REGIONAL DE BANCO DE DADOS, ERBD, 1., 2005, Porto Alegre. **Anais...** Porto Alegre: SBC, 2005. p. 55-60.

MOTRO, A. VAGUE: a user interface to relational databases that permits vague queries. **ACM Trans. Inf. Syst.**, New York, v. 6, n. 3, p. 187-214, 1988.

NAMBIAR, U.; KAMBHAMPATI, S. Answering imprecise database queries: a novel approach. In: ACM INTERNATIONAL WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 2003. **Proceedings...** New York: ACM Press, 2003. p.126-133.

NAVARRO, G. A Guided Tour to Approximate String Matching. **ACM Computing Surveys**, New York, v. 33, n. 1, p. 31-88, Mar. 2001.

ORTEGA-BINDERBERGER, M. **Integrating Similarity Based Retrieval and Query Reffinement in Databases.** 2002. PhD thesis - UIUC - University of Illinois at Urbana-Champaign, Urbana, Illinois.

RASMUSSEN, E. Clustering Algorithms. In: BAEZA-YATES, R; FRAKES, W. B. **Information Retrieval – Data Structures & Algorithms**. Rio de Janeiro: Prentice-Hall do Brasil, 1992. p. 419-442

REZAEE, R.; LELIEVELDT, B. P. F.; REIBER, J. H. C. A new cluster validity, index for the fuzzy c-mean, **Pattern Recognition Letters**, Amsterdam, v. 19, p. 237-246, 1998.

SALTON, G.; LESK, M. E. Computer Evaluation of Indexing and Text Processing. **J. ACM**, New York, v. 15, n. 1, p. 8-36, 1968.

SANTOS, R. G. **Análise e Avaliação de Algoritmos de *Clustering***. 2003. 52 f. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

STASIU, R. K. **Avaliação da qualidade de funções de similaridade no contexto de consultas por abrangência.** 2007. Tese (Doutorado em Ciência da Computação) - Instituto de Informática, (UFRGS - Universidade Federal do Rio Grande do Sul), Porto Alegre, RS.

STASIU, R. K.; HEUSER, C. A; SILVA, R. Estimating recall and precision for vague queries in databases. In: CONFERENCE ON ADVANCED INFORMATION

SYSTEMS ENGINEERING, 17., 2005, Porto, Portugal. **Advanced Information Systems Engineering**: proceedings. Berlin: Springer, 2005. p. 187-200. (Lecture Notes in Computer Science, v. 3520).

SCHALLEHN, E.; SATTLER, K.-U.; SAAKE, G. Efficient similarity-based operations for data integration. **Data Knowl. Eng.**, [S.l.], v. 48, n. 3, p. 361-387, 2004.

SNEATH, P. H.; SOKAL, R. R. **Numerical Taxonomy**. San Francisco: W. H. Freeman and Company, 1973.

SUBHASH, S. **Applied multivariate techniques**. New York: John Wiley & Sons, 1996.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2006.

THECODORIDIS, S.; KOUTZOUBAS, K. **Pattern recognition**. [S.l.]: Academic Press, 1999.

ULLMAN, J. D.; GARCIA-MOLINA, H.; WIDOM, J. **Database Systems**: the complete book. Upper Saddle River, New Jersey, USA: Prentice Hall, 2002.

WARD, J. H. J. Hierarchical grouping to optimize an objective function. **J. Am. Stat. Assoc.**, [S.l.], v. 58, p. 236-244, 1963.

WINKLER, W. E. The state of record linkage and current research problems. [S.l.]: Statistical Research Division - U.S. Bureau of the Census, 1999. (R99/04).

WIVES, L. K. **Utilizando Conceitos como descritores de Textos para o processo de identificação de conglomerados (clustering) de documentos.** 2004. 136 f. Tese (Doutorado em Ciência da Computação) - Instituto de Informática, UFRGS, Porto Alegre.

XIE, X. L.; BERFI, G. A Validity measure for Fuzzy Clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v. 13, n. 8, p. 841-847, Aug. 1991.