Faculdade de Economia da Universidade do Porto

Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Collaborative Filtering **para recomendação de produtos** ***on-line***

Filipe Fortuna

Orientador : Prof. Dr. Carlos Soares

Porto, Dezembro de 2009

À minha namorada, pais, irmão, cunhada e sobrinho

Resumo

Devido ao aumento da quantidade de informação disponibilizada através da Internet e à diminuição do tempo disponível para as pessoas procurarem o que pretendem ou necessitam, tornou-se premente encontrar formas de responder e diminuir o impacto destas duas premissas, com o intuito de aumentar a satisfação das pessoas.

Assim surgiram os Sistemas de Recomendação, com o propósito de implementar técnicas que possibilitem a sugestão de produtos ou informação considerada relevante para o utilizador do sítio WEB. Por exemplo, quando um utilizador consulta uma página relativa a um portátil, a recomendação de uma pasta para o portátil e de um rato *USB* irá eventualmente permitir poupar tempo de pesquisa ao utilizador, pois são dois produtos necessários aquando da compra de um portátil. Os Sistemas de Recomendação encontram-se amplamente disseminados pela Internet, com presença nos maiores sítios de compras *on-line* como *Amazon*, *eBay*, *Drugstore*, entre outros.

O caso de estudo considerado nesta dissertação incide sobre um sítio de comércio electrónico de produtos informáticos, o qual disponibilizou dados relativos aos *Web Access Logs* (ficheiro com todos os acessos ao sítio), produtos, subcategorias e categorias. O objectivo é avaliar comparativamente variantes da abordagem *Collaborative Filtering*, para recomendação de produtos informáticos. As variantes consideradas têm impacto ao nível do algoritmo e dos dados. A aplicação destas variantes realizou-se em duas fases distintas da geração de recomendações: quando se está a determinar quais os utilizadores do passado mais semelhantes ao utilizador que está a visitar o sítio no presente; e na fase em que se pretendem determinar os produtos que hipotéticamente são de maior interesse para o utilizador que está a visitar o sitio.

Na primeira fase, ao nível do algoritmo procedeu-se à implementação da correlação e das variantes *Default Vote* e *Inverse User Frequency*. Relativamente aos dados formularam-se três variantes: utilizar o conjunto de dados relativo aos acessos a produtos; utilizar os conjuntos de dados referentes aos acessos a produtos, subcategorias ou categorias; e utilizar informação relativa ao intervalo de tempo que o utilizador esteve a visualizar determinado produto, pois é mencionado na literatura como sendo indicador de preferência. A utilização da informação relativa ao tempo é feita através do seu valor ou através de funções utilidade, que têm como intuito modelar a importância do tempo, de acordo com o hipotético interesse dos utilizadores.

Na segunda fase, no que se refere à abordagem consideraram-se três cenários: a utilização dos acessos a produtos; a utilização do informação relativa ao intervalo de tempo em conjunto com os acessos a produtos tendo em conta situações distintas; e a utilização da informação relativa ao intervalo de tempo em conjunto com os acessos a produtos independentemente da situação. Ao nível dos dados utilizaram-se duas variantes: o conjunto de dados relativo aos acessos a produtos e os dados relativos ao intervalo de tempo de visualização de produtos.

Os resultados obtidos na primeira fase apontam para a utilização de diferentes abordagens para diferentes cenários. Não se verifica a existência de abordagens claramente superiores às restantes, portanto a escolha depende das características do Sistema de Recomendação que se pretende implementar, ou seja, do momento da recomendação (quando o utilizador já acedeu a 2, 5, ou 10 itens). Também surge outro factor a ter em conta, relacionado com a percentagem de sessões (conjunto de produtos visitados em determinado período de tempo) com recomendação, pois há abordagens que conseguem realizar recomendação em maior percentagem de sessões que outras, embora haja em determinados casos prejuízo do acerto. O implementador terá que decidir se é mais importante recomendar em menos casos, mas recomendar produtos com maior potencial de acerto nas preferências do utilizador ou pretende recomendar em mais casos prejudicando o potencial de acerto da recomendação.

Relativamente aos resultados obtidos na segunda fase, concluiu-se que há grande variabilidade nas abordagens que contribuíram para os melhores resultados. Tal como anteriormente, a abordagem a utilizar deve ter em conta as características do Sistema de Recomendação que se pretende implementar, neste caso depende se queremos recomendar 1, 3, 5 ou 10 produtos.

Agradecimentos

Gostaria de deixar o meu muito obrigado a todas as pessoas que, de algum modo, contribuíram para a realização da minha dissertação e, sendo este um momento especial na minha vida, agradecer a quem me ajudou a chegar até aqui.

Agradeço ao Prof. Dr. Carlos Soares pela orientação e acompanhamento prestado, pela partilha de conhecimento e pela disponibilidade demonstrada durante o longo percurso percorrido na realização desta dissertação.

Agradeço ao Dr. Marcos Aurélio Domingues por todo o trabalho, empenho e troca de conhecimento envolvido no tratamento e carregamento dos dados presentes nos *Web Access Logs* para o *data warehouse*, sem o qual não se poderia dar início à realização desta dissertação.

Agradeço ao Prof. Dr. Alípio Jorge pela orientação inicial que permitiu a focalização e tangibilidade dos objectivos propostos para a dissertação.

Agradeço ao Eng. Mário Machado, sócio gerente da Suprides, pela autorização concedida para fornecimento dos dados necessários para a realização desta dissertação, nos moldes actuais. Também me apraz salientar o seu interesse demonstrado na temática abordada nesta dissertação.

Agradeço ao Eng. António Pinto, administrador do sítio `WEB` de comercio electrónico da Suprides, pela colaboração na disponibilização dos dados necessários para a realização desta dissertação e por toda a colaboração e disponibilidade demonstrada.

Agradeço muito especialmente à minha namorada que me acompanhou, compreendeu, apoiou e animou ao longo desta longa caminhada. Por toda a colaboração e disponibilidade demonstrada, nomeadamente na leitura da dissertação.

Agradeço à minha família o apoio, compreensão e animo prestado para a concretização desta etapa da vida, com prejuízo da disponibilidade para estar com eles.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Devido à expansão que o acesso à Internet tem sofrido ao longo dos anos, tem aumentado a necessidade de analisar toda a informação obtida e disponibilizada através deste meio de comunicação. Na figura 1.1 pode-se constatar a percentagem de utilizadores de Internet relativamente à população mundial e o crescimento do número de utilizadores registado entre 2000 e 2008. Pode verificar-se um grande crescimento ao nível de todas as áreas do mundo, com especial ênfase no Médio Oriente, África e América Sul/Central. É de salientar que a Ásia já representa 41,2% dos utilizadores de mundiais e que apenas apresenta uma taxa de penetração de 17,4%. Como seria de esperar as regiões que englobam os países desenvolvidos, América do Norte, Oceânia/Austrália e Europa, apresentam as taxas de penetração mais elevadas.

Figura 1.1: Utilização mundial da Internet e população mundial [33].

Regiões do Mundo	População (2008 Est.)	Utilizadores de *Internet* Dez. 31, 2000	Utilizadores de *Internet* Data mais recente	Penetração (% População)	Crescimento de utilizadores 2000-2008	% de utilizadores do total
Africa	975.330.899	4.514.400	**54.171.500**	5,6 %	1.100,1 %	3,4 %
Asia	3.780.819.792	114.304.000	**657.170.816**	17,4 %	474,9 %	41,2 %
Europa	803.903.540	105.096.093	**393.373.398**	48,9 %	274,3 %	24,6 %
Médio Oriente	196.767.614	3.284.800	**45.861.346**	23,3 %	1.296,2 %	2,9 %
América do Norte	337.572.949	108.096.800	**251.290.489**	74,4 %	132,5 %	15,7 %
América Sul/Central	581.249.892	18.068.919	**173.619.140**	29,9 %	860,9 %	10,9 %
Oceânia / Austrália	34.384.384	7.620.480	**20.783.419**	60,4 %	172,7 %	1,3 %
Total no Mundo	6.710.029.070	360.985.492	**1.596.270.108**	**23,8** %	342,2 %	100,0 %

Cada vez mais a Internet disponibiliza grandes "hipermercados" *on-line*, por isso é necessário verificar de que forma os potenciais clientes que por lá navegam se comportam; tal como se realizou no passado ao nível dos hipermercados.

Analisar o comportamento dos utilizadores/clientes permite entender e identificar os seus gostos, as suas preferências, a forma como se comportam e reagem perante determinados estímulos. A identificação do tipo de utilizadores/clientes de uma loja *on-line* constitui, cada vez mais, uma preocupação das empresas que disponibilizam serviços ou produtos por esta via. Esta atitude tem como objectivo inerente melhorar o acesso à informação e aos produtos/serviços desejados pelo cliente. Não

é do interesse destas empresas que o cliente perca muito tempo ou ande perdido à deriva para encontrar o que pretende, de forma a que não perca o interesse e desista. Por outro lado, devem proporcionar ao utilizador publicidade direccionada aos seus interesses, de forma a aumentar as hipóteses de venda e satisfação do utilizador. Em suma, num mercado globalizado como o actual, uma empresa deve conhecer muito bem os seus clientes, de forma a aumentar o grau de satisfação destes, verificando-se uma adaptação da empresa às necessidades do cliente, o que tipicamente proporciona um aumento das vendas da empresa.

O que foi dito no paragrafo anterior pode ser sintetizado pela afirmação de Jeff Bezos, director executivo da *Amazon*: "Se houvesse 3 milhões de consumidores na `WEB`, deveria haver 3 milhões de lojas na `WEB`". O grau de satisfação assume grande importância uma vez que segundo a revista *Nielsen* [34] os utilizadores *on-line* realizam as suas compras nos sítios em que estão familiarizados e 60% afirma que utilizam quase sempre o mesmo sítio [34]. O comercio electrónico tem registado um grande crescimento, pois 85% da população *on-line* já utilizou a `Internet` para realizar uma compra e verificou-se um aumento de 40% das compras *on-line* em 2006 e 2007 [34]. Os países com maior percentagem de utilizadores que realizam compras *on-line* são: Coreia do Sul (99%), Reino Unido (97%), Alemanha (97%) e Japão (97%), com os Estados Unidos a aparecerem em oitavo lugar com 94%. Globalmente os itens mais comprados através da `Internet` são [34]:

- Livros (41%);
- Roupas, Acessórios, Sapatos (24%);
- Vídeos, DVDs, Jogos (24%);
- Bilhetes para companhias aéreas (24%);
- Equipamento Electrónico (23%).

Perante esta realidade, o estudo de `Sistemas de Recomendação`, com o propósito de implementar técnicas que possibilitem a sugestão de itens (informação, produtos ou serviços) considerada relevante para o utilizador do sítio `WEB`, tem também ganho cada vez mais relevância. A geração das sugestões pode ser realizada a partir de diferentes fontes de informação, métodos de recolha da informação e abordagens para obtenção das recomendações.

Como fontes de informação temos: dados demográficos, características dos itens e preferências dos utilizadores.

Os métodos de recolha da informação podem essencialmente caracterizar-se em: explicita ou implícita. No primeiro tipo o utilizador do sítio de `Internet` declara explicitamente a avaliação que faz de determinado item. Por exemplo no sítio *MovieLens* o utilizador classifica os filmes de 1 (Mau) a 5 (Muito Bom), consoante os seus gostos. No segundo tipo utiliza-se informação relativa ao comportamento e escolhas do utilizador ao longo da sua interacção com o sítio. Esta informação pode ser retirada dos *Web Access Logs* (ficheiros onde ficam registados todos os acessos ao sítio), do histórico de compras dos utilizadores, da impressão de determinada página ou do facto de uma página ser adicionada aos favoritos [6]. No caso de estudo utilizou-se o método implícito, pois é o método indicado para as fontes de informação existentes.

Relativamente às abordagens, existem sensivelmente três grandes áreas de estudo, sistemas baseados no conteúdo, *Collaborative Filtering* e sistemas híbridos.

Os sistemas baseados no conteúdo realizam a recomendação de itens, tendo em conta as características (marca, tipo, modelo, tamanho, autor, etc.) dos itens considerados como relevantes pelo utilizador no passado [4]. Por exemplo se um utilizador viu vários portáteis de 12,5 polegadas da *Toshiba*, poderá significar que este utilizador está interessado em produtos da marca *Toshiba*, de tamanho de 12,5 polegadas e do tipo portáteis, portanto devem ser sugeridos produtos com estas características. Do mesmo modo se o utilizador visualizar uma série de monitores de 15 polegadas de várias marcas, poderá significar que a marca não é o mais relevante mas sim o tamanho.

O sistema *Collaborative Filtering* caracteriza-se pela recomendação de itens a partir do histórico de itens consultados ou avaliados por outros utilizadores. Sendo assim, as recomendações são geradas com base na similaridade de utilizadores, a partir dos itens já visualizados pelo utilizador actual, são calculados os utilizadores com preferências mais semelhantes e recomendados os itens mais relevantes que o utilizador actual ainda não consultou [4]. Por exemplo, se um utilizador acede a dois portáteis de marca *Asus* de modelos diferentes, por comparação com outros utilizadores que viram estes dois itens no passado, permite obter a sugestão do portátil de marca *INMOVE*. Do mesmo modo para um utilizador que acede a um processador e uma memória, origina a sugestão de uma placa gráfica.

Os sistemas híbridos resultam da junção das duas abordagens anteriores, constituindo a forma como se realiza a junção, diferentes variantes desta abordagem. Surgiram com o propósito de aproveitar as vantagens das duas abordagens anteriores e colmatar as limitações de cada uma delas isoladamente. Por exemplo, utilizar as técnicas tradicionais dos sistemas de *Collaborative Filtering*, mas mantendo informação relativa às características dos itens visitados no passado pelo utilizador, pois esta será utilizada para determinar a semelhança entre utilizadores, em vez da solução tradicional em que a semelhança é determinada apenas com base nos itens relevantes comuns [12].

Esta dissertação tem como objectivo avaliar comparativamente variantes da abordagem *Collaborative Filtering*, para recomendação de itens. As variantes consideradas têm impacto ao nível do algoritmo e dos dados. A aplicação destas variantes realizou-se em duas fases distintas da geração de recomendações:

- quando se está a determinar quais os utilizadores do passado mais semelhantes ao utilizador que está a visitar o sítio no presente:
  - implementar a correlação e as variantes *Default Vote* e *Inverse User Frequency* utilizando o conjunto de dados dos acessos a itens;
  - utilizar os conjuntos de dados referentes aos acessos a itens, subcategorias ou categorias;
  - utilizar informação relativa ao intervalo de tempo que o utilizador esteve a visualizar determinado item.
- quando se pretendem determinar os itens que hipotéticamente são de maior interesse para o utilizador que está a visitar o sitio:
  - utilizar os acessos a itens;
  - utilizar a informação relativa ao intervalo de tempo em conjunto com os acessos a itens tendo em conta condicionantes distintas;

– utilizar a informação relativa ao intervalo de tempo em conjunto com os acessos a itens independentemente das condicionantes.

Na primeira fase procedeu-se à implementação de uma das variantes dos sistemas de *Collaborative Filtering*, mais concretamente os algoritmos *Memory-based* que se caracterizam pelo facto de utilizarem dados que traduzem preferências dos utilizadores no passado para realizarem recomendações. Dentro deste tipo de algoritmo foi utilizada a Correlação de *Pearson* que tem como intuito descortinar a semelhança entre utilizadores. Para além deste foram ainda implementadas duas variantes dos algoritmos *Memory-based* que são: *Default Vote* e *Inverse User Frequency*. O primeiro é especialmente indicado para situações em que os dados são muito esparsos e o segundo visa determinar a importância de determinado item na obtenção da semelhança entre utilizadores, através da frequência, isto é, se dois utilizadores visitarem um item com poucos acessos, diz mais da semelhança entre eles que um item que tem mais acessos.

Para além das variantes dos algoritmos procedeu-se também à variação da informação a partir da qual se realizaram as recomendações. Foram considerados dados relativos aos *Web Access Logs*, subcategorias e categorias dos produtos e dados relativos ao intervalo de tempo que determinado utilizador passou na visualização de um item. A integração destes diferentes tipos de dados têm como intuito obter mais informação acerca dos utilizadores de forma a realizar recomendações o mais próximas possível das suas preferência.

Na segunda fase realizou-se a recomendação dos itens de maior valor previsto, pelo que as variantes utilizadas visam implementar estratégias que hipotéticamente melhoram a identificação dos itens nos quais os utilizadores estão interessados. Também nesta fase se utiliza a integração de informação, nomeadamente informação relativa ao tempo, com o mesmo propósito apresentado para a primeira fase.

As variantes mencionadas anteriormente foram avaliadas comparativamente, quer ao nível dos algoritmos, quer ao nível dos dados, de forma a determinar o impacto que cada variante tem sobre a fiabilidade das recomendações geradas.

Os dados utilizados são provenientes de um sítio de comércio electrónico de produtos informáticos, com um número total de visitantes em 2006 de 105072 e um total de 69,9 *GigaBytes* de tráfego.

1.0.1 Estrutura da dissertação

Esta dissertação encontra-se organizada em sete capítulos, incluindo a Introdução, Conclusão, Bibliografia e Anexos.

O capítulo denominado de "Sistemas de Recomendação" visa apresentar qual o propósito e origem dos sistemas de recomendação, assim como os diferentes tipos de sistemas existentes e diferentes abordagens. Mais concretamente será apresentada a abordagem de *Collaborative Filtering* e respectiva organização dos dados, algoritmos (*Memory-Based*: correlação, semelhança vectorial e variantes; *Model-Based*: Cluster models, Bayesian networks e Regras de associação) e limitações. Também são apresentadas as abordagens baseada no conteúdo e híbrida, com as respectivas limitações. É ainda referido qual o procedimento de avaliação e quais as métricas a utilizar; são mencionados alguns exemplos de sistemas de recomendação quer na vertente académica, quer na vertente comercial; e por

fim são apresentados os desafios dos sistemas.

O capítulo seguinte "Enriquecimento dos dados para *Collaborative Filtering*" caracteriza-se pela apresentação de duas contribuições, com o propósito de integração de informação adicional: tempo, categorias e subcategorias. Para melhor enquadrar as contribuições realiza-se ainda a apresentação da empresa, dos dados fornecidos, pré-processamento, organização dos dados utilizados no sistema de recomendação e por fim análise exploratória dos dados. Com esta última pretende-se determinar quais os itens, categorias e subcategorias mais visitadas; percentagem de visitas a itens e subcategorias em sessões distintas; e percentagem de visitas a itens, categorias e subcategorias , por sessão.

O capítulo "Experiências e Resultados" visa apresentar as metodologias de experimentação, os resultados, análise e discussão de resultados.

Ao nível dos anexos temos mais dois capítulos com o propósito de apresentar com maior detalhe as ferramentas de apoio e o trabalho realizado no âmbito desta dissertação. No primeiro capítulo descreve-se a organização dos dados no *data warehouse*, cuja origem são os *Web Access Logs* e a obtenção e tratamento dos dados necessários para o sistema de recomendação. No segundo capítulo são apresentados detalhes técnicos da implementação do sistema de recomendação.

Capítulo 2

Sistemas de Recomendação

Este capítulo visa apresentar a importância e o papel dos Sistemas de Recomendação na Internet, especialmente ao nível do Comercio Electrónico. São descritas as características dos Sistemas de Recomendação, assim como as fases, variantes e limitações das várias abordagens de geração de recomendações para o utilizador (sistemas baseados em *Collaborative Filtering*, conteúdo e híbridos), com especial atenção para os sistemas baseados em *Collaborative Filtering*, pois são objecto de estudo nesta dissertação. Também são disponibilizados exemplos de Sistemas de Recomendação, de origem académica e de origem comercial, com o intuito de proporcionar a concretização dos conceitos apresentados. No final deste capítulo são apresentados os desafios que se colocam para a evolução dos Sistemas de Recomendação.

2.1 Contexto e Características dos Sistemas de Recomendação

Na actualidade o tempo é um recurso escasso que todos querem utilizar da forma o mais proveitosa possível. Como tal, os sítios WEB que possibilitam reduzir o tempo de procura do que se pretende, tendem a ser cada vez mais valorizados e apreciados. Por outro lado, temos sítios WEB cada vez mais elaborados e com quantidades de produtos e serviços cada vez maiores, aumentando sem dúvida a oferta, a diversidade e a capacidade de satisfazer as necessidades de quem procura, no entanto estas mais valias conduzem a uma maior dificuldade na procura do que se pretende e consequentemente à perda de tempo.

Tais factos têm impulsionado os Sistemas de Recomendação, pois estes têm como intuito facilitar o acesso dos utilizadores/clientes à informação, produtos e serviços. Uma maior facilidade de acesso permite que o utilizador/cliente perca menos tempo na pesquisa do que pretende, pelo que se potencia a satisfação e aumentam as vendas. Ao sugerir informação e produtos de interesse aos utilizadores/clientes facilitam-se as relações e personalizam-se os serviços.

Os principais objectivos dos Sistemas de Recomendação são [4]:

- Converter utilizadores em compradores;
- Melhorar o *design* do sítio WEB;
- Aumentar a lealdade e a retenção de clientes;

- Aumentar o *cross-sell* através da recomendação de itens relacionados com os que o utilizador está a consultar;
- Ajudar os utilizadores a encontrar mais rapidamente o que pretendem;
- Permitir que a oferta de produtos/serviços esteja mais relacionada com o contexto e interesses do utilizador.

Tendo em conta o caso de estudo, é disso exemplo a situação em que o utilizador pretende adquirir um portátil, tipicamente também necessita de um rato *USB* e uma mala. Num sítio sem `Sistemas de Recomendação` o utilizador terá que despender tempo a procurar estes itens. Com os SR estes itens podem ser sugeridos automaticamente, evitando o tempo despendido na procura e possibilitando também um reforço positivo para o comprador, pois é-lhe proporcionado conhecimento relativo ao que outros utilizadores compraram ou viram em conjunto com o portátil.

2.1.1 Fontes de informação

Designam-se por fontes de informação todas as origens de onde provêem os dados utilizados para realizar determinado estudo. Para levar a cabo a implementação de `Sistemas de Recomendação` podem ser utilizadas diferentes fontes de informação. Uma destas fontes de informação são os dados sócio-demográficos como idade, sexo, ocupação, rendimentos, níveis de instrução, *hobbies*. Através destes dados e dos itens adquiridos ou preferenciais, podem-se inferir relações e agrupar utilizadores/clientes[5]. Por exemplo, os adolescentes preferem itens que chamem atenção, enquanto que as pessoas que se enquadram na faixa etária dos 30 aos 40 preferem artigos mais sóbrios.

Uma outra fonte de informação são as características dos itens, que podem ser: extrínsecas ou intrínsecas. Como características extrínsecas temos a cor, tipo, fabricante, categoria. Os interesses dos utilizadores podem ser obtidos através de estudos de mercado. Por exemplo há determinadas marcas que inspiram mais confiança aos clientes que outras. As características intrínsecas são obtidas através da análise do conteúdo dos itens, isto é, como os consumidores e o próprio fabricante definem o produto relativamente a conceitos que avaliam o produto como um todo, por exemplo a portabilidade, a flexibilidade, a resistência, a durabilidade, etc..

O terceiro tipo de informação utilizada nos `Sistemas de Recomendação` são as preferências dos utilizadores, que podem ser obtidas através de avaliação explícita e implícita. A avaliação explícita consiste na execução de inquéritos aos utilizadores onde lhes é pedido para identificarem os seus interesses e os classificarem segundo um *ranking*, utilizando métricas binárias ou escalares. Este tipo de metodologia tem como limitações o facto de por vezes se verificar a classificação dos itens de forma desinteressada, o que pode conduzir a classificações erróneas; nem toda a informação consultada é classificada; e contribui para uma maior incidência de desistências da utilização dos sítios. A avaliação implícita está relacionada com toda a informação que se pode obter acerca dos interesses de um determinado utilizador de forma não intrusiva (sem participação directa do utilizador) como exemplos desta informação temos os *Web Access Logs* dos servidores `WEB`, histórico de compras, padrões de navegação, colocar páginas nos favoritos, imprimir páginas, intervalo de tempo despendido pelo utilizador na visualização de determina página, etc.. Este tipo de avaliação apresenta menor precisão que a avaliação explicita, mas existem estudos que indicam que os factores implícitos como

a utilização de *Web Access Logs* e o tempo despendido numa página estão fortemente correlacionados com a avaliação explícita [7], isto é, estes factores contribuem para a melhoria das recomendações obtidas de uma forma implícita.

2.1.2 Tipos de decisões de recomendação

Para implementação de `Sistemas de Recomendação` tem que se ter em conta uma das seguintes decisões [5]:

- Previsão ⇒ prevê as preferências dos utilizadores para itens que este ainda não avaliou ou escolheu.
- Top-N recomendações ⇒ lista de N itens que o utilizador activo poderá gostar, mas que ainda não escolheu ou avaliou.
- Top-M utilizadores ⇒ Quando há novos itens o sistema sugere-os aos utilizadores que potencialmente têm maior preferência por estes itens.

No primeiro ponto temos os `SR` que explicitam o valor da sua previsão através de uma escala, permitindo ao utilizador ficar com uma ideia acerca do interesse que o item suscita, como o sistema *MovieLens*, onde os utilizadores podem seleccionar filmes que ainda não viram e requer que seja gerada uma previsão numa escala de uma (não gosta) a cinco estrelas (adora) (ver 2.5.1).
O segundo ponto corresponde aos `SR` que fornecem ao utilizador uma lista de itens, como o sítio de comércio electrónico *Amazon.com*, onde são apresentadas ao utilizador as listas dos itens que foram comprados por utilizadores que também compraram este item (ver 2.5.2).
O terceiro ponto engloba os `SR` que recomendam itens com características semelhantes às dos itens que o utilizador gostou no passado. No sítio da *Amazon.com* depois do utilizador se autenticar são lhe sugeridos itens com características semelhantes aos itens que este comprou no passado.

2.1.3 Tipos de `Sistemas de Recomendação`

Em seguida vai proceder-se à apresentação dos `SR` que realizam recomendações personalizadas, ou seja, sugestão de produtos/serviços baseados nos itens já vistos no passado pelo utilizador ou perfil do utilizador resultante de avaliação explícita. Estes distinguem-se consoante a fonte de informação e a técnica utilizada, sendo os mais os frequentes classificados da seguinte forma:

- Baseada no Conteúdo ⇒ recomendação de itens semelhantes aos preferidos no passado.
- *Collaborative Filtering* ⇒ identificação de utilizadores com preferências similares.
- Híbridos ⇒ são sistemas que juntam as duas abordagens anteriores com o intuito de colmatar os pontos fracos de cada uma delas.

Para além dos `SR` mencionados anteriormente, ainda existem outros sistemas menos utilizados como os baseados em Associação que se caracterizam pela identificação de itens frequentemente associados a itens escolhidos pelos utilizadores no passado, em Demografia que geram recomendações baseadas

nas preferências dos utilizadores de acordo com as características demográficas e em Reputação onde ocorre a identificação de utilizadores nos quais um utilizador se revê e utilizam-se as opções destes utilizadores para realizar as recomendações. Esta é uma nova abordagem baseada nas ciências sociais [5].

Os sistemas baseados na Popularidade são `SR` que não se baseiam na personalização, porque se caracterizam pela recomendação de itens tendo em conta a percentagem de utilizadores que compraram o item e/ou tendo em conta a média da avaliação atribuída pelos utilizadores [5]. Normalmente os utilizadores gostam de ter noção dos produtos mais vendidos e com melhor avaliação. Devido ao facto de serem fáceis de implementar, têm uma ampla difusão pela `Internet`.

2.2 *Collaborative Filtering*

Nesta secção pretende-se apresentar algumas variantes de `Sistemas de Recomendação` baseados em *Collaborative Filtering*.

2.2.1 Organização dos dados

Nos sistemas de *Collaborative Filtering* os conjuntos de dados (dados utilizados pelos algoritmos para realizar as recomendações/previsões) são tipicamente representados por uma matriz onde as colunas representam itens e as linhas utilizadores. Na intersecção destas são identificadas as preferências dos utilizadores relativamente aos itens ($v_{i,j}$). A informação que permite inferir as preferências pode ter uma ou várias origens.

	j_1	j_2	..	j_l	..	j_k
i_1						
i_2		$v_{2,2}$				
:						
i_a						
i_q				$v_{q,l}$		
i_m						

Tabela 2.1: Esquema da organização conceptual do conjunto de dados.

Como se pode observar através da tabela 2.1 temos m utilizadores $U = \{i_1, i_2, .., i_m\}$, que podem estar interessados em k itens $I = \{j_1, j_2, .., j_k\}$. Cada utilizador tem uma lista de itens I_i, pelos quais tem preferência, isto é, itens pelos quais o utilizador demonstrou, interesse no passado, este interesse é determinado de forma explicita através da avaliação dos itens ou de forma implícita através de *Web Access Logs*, histórico das compras, se imprimiu, se colocou nos favoritos, o intervalo de tempo de visualização do item, etc.. No esquema é também apresentada a sigla i_a que corresponde ao utilizador activo (utilizador que se encontram presentemente a aceder aos itens), para o qual se pretende prever a preferência para determinado item ou realizar a recomendação de um ou mais itens pelos quais o utilizador activo estará hipotéticamente interessado.

2.2.2 Algoritmos *Memory-Based*

Os algoritmos *Memory-Based* caracterizam-se pela utilização de uma base de dados com o histórico da avaliação (votos) dos itens levada a cabo pelos utilizadores, a partir da qual se procuram prever recomendações para os utilizadores activos do sítio. O cálculo da média dos votos de um dado utilizador *i* recorre ao conjunto de itens em que *i* votou (I_i) e aos votos que o utilizador *i* fez no item *j* ($v_{i,j}$). A equação é a seguinte [6]:

$$\overline{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

Esta média será utilizada para a determinação do voto previsto do utilizador activo para o item *j*. O $p_{a,j}$ é obtido através da formula [6]:

$$p_{a,j} = \overline{v}_a + k \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v}_i) \tag{2.1}$$

O *n* é número de utilizadores existentes na base de dados que votaram no item *j*. O peso $w(a,i)$ reflecte a distância, correlação ou a similaridade entre os utilizadores *i* e o utilizador activo. O peso visa quantificar a semelhança entre os utilizadores existentes na base de dados e o utilizador activo tendo em conta os votos. Por fim, o *k* é um factor de normalização.

Correlação

Para cálculo do peso será utilizada a correlação de *Pearson* que é caracterizada pela seguinte formula [6]:

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}} \tag{2.2}$$

O somatório de *j* corresponde aos itens em que ambos os utilizadores votaram.

Semelhança Vectorial

A Semelhança Vectorial é oriunda da área de "Information retrieval" e pode também ser utilizada para cálculo do peso w. Este cálculo permite medir a semelhança entre dois documentos. Para tal representa-os como vectores da frequência das palavras que constituem o documento e calcula o coseno do ângulo formado pelos dois vectores de frequências [6].

A adaptação desta medida para a área de *Collaborative Filtering* considera os utilizadores como documentos e os itens como palavras e os votos como a frequência das palavras. A equação que traduz este algoritmo é [6]:

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}} \tag{2.3}$$

Os termos das raízes quadradas têm como intuito normalizar os votos, para que os utilizadores que votam em muitos itens não sejam *a priori* beneficiados relativamente aos restantes utilizadores. Também são possíveis outros esquemas de normalização como a soma absoluta e número de votos.

Variantes dos algoritmos *Memory-Based*

Nesta secção são apresentadas algumas variantes dos algoritmos *Memory-Based* para serem aplicadas em situações especificas.

Default Vote

O *Default Vote* é uma variante da medida de correlação, é especialmente indicado para casos em que existem poucos itens visitados no histórico quer ao nível do utilizador activo, quer ao nível dos utilizadores do histórico. Nestes casos a medida de correlação não apresenta uma boa performance porque utiliza apenas os itens visitados por ambos os utilizadores ($I_a \cap I_j$). Se utilizarmos alguns valores por defeito em que nenhum dos utilizadores votou também encontramos preferências coincidentes. A formula do *Default Vote* é [6]:

$$w(a,i) = \frac{(n+k)(\sum_j v_{a,j}v_{i,j} + kd^2) - (\sum_j v_{a,j} + kd)(\sum v_{i,j} + kd)}{\sqrt{((n+k)\sum_j(v_{a,j}^2 + kd^2) - \sum v_{a,j} + kd)^2)((n+k)(\sum_j v_{i,j}^2 + kd^2) - (\sum_j v_{i,j} + kd)^2)}} \tag{2.4}$$

O d deve reflectir um valor neutro ou uma preferência negativa. O primeiro caso utiliza-se quando se considera que o facto de o utilizador não visitar um item, não é a expressão de gosto, nem repudio pelo item, enquanto no segundo caso considera-se que o facto de o utilizador não visitar um item é sinal de repudio pelo item. No caso de estudo o d assume o valor de 0, com o intuito de expressar que o item não foi visitado. O somatório j incide sobre os itens que ambos os utilizadores visitaram ($I_a \cap I_j$) e $n = |I_a \cap I_j|$. Uma vez que d é 0 temos:

$$w(a,i) = \frac{(n+k)(\sum_j v_{a,j}v_{i,j}) - (\sum_j v_{a,j})(\sum v_{i,j})}{\sqrt{((n+k)\sum_j v_{a,j}^2 - (\sum v_{a,j})^2)((n+k)\sum_j v_{i,j}^2 - (\sum_j v_{i,j})^2)}} \tag{2.5}$$

Inverse User Frequency

A *Inverse User Frequency* é utilizada em aplicações de "Information retrieval" em vectores de semelhança. Tem como objectivo reduzir o peso das palavras utilizadas com maior frequência no texto, com base no pressuposto de que não são tão úteis para a identificação do assunto do documento. Esta ideia pode ser transposta para *Collaborative Filtering*, considerando que os itens visitados com maior frequência têm menos peso no calculo da semelhança entre utilizadores que os itens menos visitados. Por exemplo se um tipo de portátil é visitado por poucos utilizadores verifica-se que este facto proporciona mais informação acerca dos gostos destes utilizadores do que um que seja visto por muitos utilizadores, pois constitui a expressão de gostos específicos. A formula é, $f_j = \log \frac{n}{n_j}$, onde n_j corresponde ao número de utilizadores que votaram no item j e n é o número total de utilizadores na base de dados. A utilização da *Inverse User Frequency* no caso do algoritmo de *Semelhança Vectorial* implica a multiplicação de f_j pelo voto original [6]. A aplicação desta variante da medida da *Correlação* implica a alteração da equação 2.2 para a seguinte forma [6]:

$$w(a,i) = \frac{\sum_j f_j \sum_j f_j v_{a,j}v_{i,j} - (\sum_j f_j v_{a,j})(\sum_j f_j v_{i,j})}{\sqrt{UV}} \tag{2.6}$$

onde:

$$U = \sum_j f_j (\sum_j f_j v_{a,j}^2 - (\sum_j f_j v_{a,j})^2) \quad V = \sum_j f_j (\sum_j f_j v_{i,j}^2 - (\sum_j f_j v_{i,j})^2) \tag{2.7}$$

2.2.3 Algoritmos *Model-Based*

Os algoritmos *Model-Based* são caracterizados pela utilização do histórico dos votos dos utilizadores em determinados itens para criação de modelos que são posteriormente utilizados na recomendação. Enquanto que os algoritmos *Collaborative Filtering* utilizam o histórico dos votos dos utilizadores no momento da recomendação, de forma a identificar quais os utilizadores mais semelhantes ao utilizador activo e proceder à recomendação dos itens que o utilizador activo ainda não viu. No primeiro caso o histórico dos votos é utilizado num momento anterior à recomendação, enquanto que no segundo caso o histórico dos votos é utilizado quando se pretende realizar a recomendação.

Para o utilizador activo pretende-se prever a votação para itens ainda não visualizados. Ao assumir que os votos têm um valor inteiro com um intervalo de 0 a m temos [6]:

$$p_{a,j} = E(v_{a,j}) = \sum_{i=0}^{m} P(v_{a,j} = i | v_{a,k}, k \in I_a) i \tag{2.8}$$

P é a probabilidade de o utilizador activo realizar determinado voto no item j, dados os votos previamente observados. Dois dos modelos que se podem utilizar nesta abordagem são: *Cluster models* e *Bayesian networks*.

Cluster models

Com esta abordagem pretende-se descobrir conjuntos de utilizadores com preferências e gostos semelhantes. Este modelo pressupõe que a probabilidade dos votos é independente para determinado conjunto de dados e é lhe atribuída uma classe variável não observada C, tendo em conta relativamente poucos valores. Pode caracterizar-se pela seguinte equação [6]:

$$P(C = c, v_1, .., v_n) = P(C = c) \prod_{i=1}^{n} P(v_i | C = c) \tag{2.9}$$

Do lado esquerdo da equação temos a probabilidade de observar um individuo de uma classe e um conjunto completo de votos. Do lado direito estão identificadas as probabilidades de pertencer à classe P(C = c) e a probabilidade condicional de realizar determinado voto, dado determinada classe P($v_i|C = c$), estas são estimadas a partir do conjunto de treino, com os votos dos utilizadores. Por exemplo, há utilizadores que estão interessados em determinadas marcas ou características dos computadores portáteis, pelo que este modelo vai identificar e agrupar os utilizadores que gostam da mesma marca ou que gostam de determinada característica. A partir do momento que o utilizador activo realiza alguns votos será determinado o grupo em que se enquadra e geradas recomendações com itens da mesma marca ou com a mesma característica.

Este algoritmo tipicamente produz recomendações menos personalizadas e com pior exactidão que os outros algoritmos, embora apresente uma melhor performance em termos de rapidez, uma vez que o tamanho do conjunto de dados a analisar para gerar recomendações é muito menor, pois cinge-se ao grupo ao qual o utilizador se enquadra [10].

Bayesian networks

A utilização do algoritmo *Bayesian networks* pressupõe que os nodos correspondem aos diferentes itens. Os estados de cada nó correspondem aos votos possíveis para cada item, considerando também a possibilidade de não haver voto (para englobar os itens sem votos). O algoritmo de aprendizagem procura os vários modelos de dependências (relações) para cada item, determinando a probabilidade de um evento ocorrer dado que existe um evento anterior (probabilidades condicionais). Na rede obtida cada item tem um conjunto de pais que correspondem aos eventos que ocorreram anteriormente [6].

Os resultados obtidos para cada nodo podem ser representados através de uma árvores de decisão, que incorpora as probabilidades condicionais para o nodo. Um exemplo deste tipo de árvore é apresentado na figura 2.1 e representa um caso em que se pretende analisar a visualização de ratos da marca *Toshiba*, tem como pais "Rato USB" e "Portátil Toshiba". Da análise de um dos ramos da árvore verifica-se que se o utilizador visualizar um portátil *Toshiba* e um rato USB então vai ver um rato *Toshiba* com maior probabilidade.

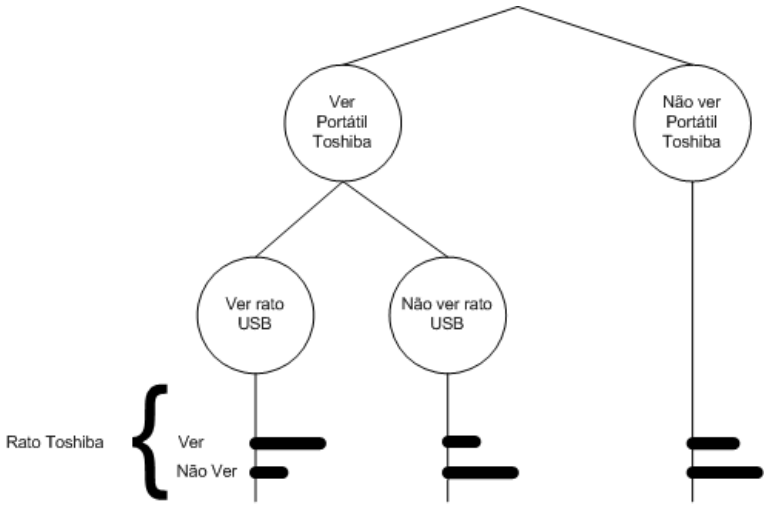


Figura 2.1: Árvore de decisão que incorpora as probabilidades condicionadas para determinado nó (exemplo meramente ilustrativo e hipotético).

A utilização deste algoritmo demonstrou bons resultados em ambientes onde as preferências mudam lentamente, tendo em conta o tempo necessário para construir o modelo. Não é adequando para ambientes em que as preferências dos utilizadores mudam rapidamente, pois a actualização frequente e em curtos espaços de tempo, conduz a muito tempo despendido quer na recolha e pré-processamento de dados (tratamento dos dados), quer na actualização posterior dos modelos [10].

Regras de associação

A utilização deste algoritmo tem como propósito encontrar várias co-ocorrências de itens num conjunto de transacções. Cada transacção corresponde a um utilizador e contém um conjunto de itens que o utilizador prefere ou comprou. Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, que representa a noção de que quando um item X é visitado, outro conjunto de itens Y frequentemente

também o são. Nesta implicação: $X \subset I, Y \subset I$ e $X \cap Y = \oslash$. Em D (conjunto de transacções) a regra $X \Rightarrow Y$ tem uma confiança c se $c\%$ das transacções em D que contêm X também contêm Y. A regra $X \Rightarrow Y$ tem um suporte s em D, se $s\%$ das transacções em D contêm $X \cup Y$. As regras de associação obtidas têm que satisfazer minimos de suporte e confiança. Tipicamente, para satisfazer as condições referidas anteriormente é utilizado o algoritmo *Apriori* [5].

No caso das regras incidirem sobre dados binário, o item j só está incluído na transacção do utilizador i se p_{ij} é 1. Se a preferência dos utilizadores estiver numa escala numérica, a decisão de inclusão de um item na transacção de determinado utilizador, pode ser feita através de um limite, baseado na média ou outros métodos. Assim, para dado limite α, um item j será incluído na transacção do utilizador i se $p_{ij} \geq \alpha$. De forma semelhante, mas utilizando um limite baseado na média, um item j será incluído numa transacção do utilizador i se $p_{ij} \geq \overline{p}_i$, onde $\overline{p}_i$ é a média das preferências de i [5].

Para levar a cabo uma recomendação, por exemplo *top-N*, para determinado utilizador é necessário descobrir as regras que são suportadas pelo utilizador, ou seja, todas as regras que apresentam itens pertencentes ao conjunto de transacções do utilizador ($X \in D_i$). Considerando os itens presentes no lado direito das regras seleccionadas anteriormente (Y) e que não aparecem no conjunto de transacções do utilizador, vai se proceder há ordenação destes itens (Y) tendo em conta a confiança das regras seleccionadas [5].

Em seguida são apresentados exemplos de regras hipotéticamente obtidas para um suporte mínimo de 1% e confiança minima de 50%. As regras são apresentadas sobre a forma de $X \Rightarrow Y|(c, s)$, onde c é confiança e s o suporte (valores em percentagem). A terceira regra diz-nos que se o utilizador vê um processador e uma memória, também vê uma placa gráfica, com 71,6% de confiança e 11,81% de suporte.

[Portátil] ⇒ [Pasta para portátil] (98,80; 5,79)
[Portátil] ⇒ [Rato USB] (79,51; 11,81)
[Processador], [Memória] ⇒ [Placa Gráfica] (71,60; 11,81)

Um dos problemas que esta técnica está susceptível é caracterizado pelo facto de as regras serem constituídas por itens com características similares (*synonymy problem*) [5].

2.2.4 Limitações dos sistemas *Collaborative Filtering*

As limitações apontadas para esta abordagem são [12]:

- Novo utilizador;
- Novo item;
- *Sparsity*.

A primeira limitação está relacionada com a dificuldade inerente à realização de recomendações a novos utilizadores, isto é, utilizadores para os quais ainda não existe histórico de itens visitados/avaliados, nestes casos o sistema não consegue realizar recomendações personalizadas. Só depois

do utilizador visitar/avaliar alguns itens é que o sistema consegue realizar recomendações, pois já se tem informação que permite encontrar utilizadores com interesses semelhantes.

No que diz respeito à segunda limitação, pode dizer-se que está relacionada com a dificuldade inerente à recomendação de itens adicionados recentemente. Isto acontece devido ao facto de não existirem quaisquer visitas/avaliações para estes itens.

A terceira limitação refere-se ao problema do número diminuto de avaliações/visitas existentes, quando comparadas com o número de itens disponíveis. Esta limitação é amplificada quando o utilizador activo tem gostos muito específicos, o que dificulta a procura de utilizadores com preferências semelhantes, afectando a qualidade das recomendações.

2.3 Outras Abordagens

Nesta secção vai-se proceder à apresentação de duas abordagens utilizadas nos `Sistemas de Recomendação` com ampla aceitação e utilização ao nível da `Internet` e mundo académico.

2.3.1 Abordagem baseada no Conteúdo

O objectivo desta abordagem é recomendar itens tendo em conta a semelhança com itens adquiridos ou vistos no passado. A semelhança é determinada de acordo com as características dos itens, como por exemplo a marca, potencia, velocidade. Apresenta melhores resultados na inclusão no `Sistemas de Recomendação` de novos itens, pois não necessita de avaliação ou compras anteriores para gerar recomendações [5].

Formalmente os SR baseados no conteúdo podem ser caracterizados pela utilização da utilidade $ut(i, j)$, onde j é o item estimado para o utilizador i, baseado na utilidade $ut(i, j_l)$ atribuída pelo utilizador i aos itens $j_l \in I$, que é semelhante ao item j. Por exemplo, para recomendar filmes ao utilizador i, o SR baseado no conteúdo tenta estabelecer/detectar as semelhanças entre os filmes que o utilizador i avaliou com classificação mais elevada no passado (por exemplo através do actor, directores, géneros, temas, etc.). Só os filmes com maior grau de semelhança com as preferências do utilizador são recomendados [12].

Para a obtenção das recomendações é necessário que haja uma fase de aprendizagem do perfil do utilizador que visa estabelecer a relação entre a avaliação do utilizador e as características do item avaliado. Para realizar esta tarefa podem ser utilizadas técnicas da estatística (especialmente regressão linear múltipla), aprendizagem indutiva (especialmente a árvore `CART` (Classification and Regression Trees)) e probabilidades *Bayesian* [5].

A regressão linear múltipla baseia-se no pressuposto de que há uma influência linear de cada característica que envolve a preferência. A equação utilizada é:

$$p_{i,j} = \sum_{c=1}^{k} w_c f_{j,c} + b \tag{2.10}$$

Onde $p_{i,j}$ é a preferência do utilizador i no item j; w_c é o coeficiente associado com a característica c; $f_{j,c}$ é o valor da característica c do item j; e b representa o enviesamento.

Ao nível da aprendizagem indutiva a avaliação realizada pelo utilizador é considerada a classe (classifica os atributos) e as características do item os atributos (constituintes de um objecto). As árvores

de decisão caracterizam-se pela utilização da estratégia de "dividir para conquistar", onde um problema complexo é decomposto em sub-problemas mais simples e posteriormente a mesma estratégia é aplicada aos sub-problemas. A capacidade de discriminação de uma árvore tem origem na divisão do espaço definido pelos atributos em sub-espaços; a cada sub-espaço é associada uma classe. Na representação de árvores cada nodo de decisão contém um teste num atributo; cada ramo descendente corresponde a um possível valor desse atributo; Cada folha está associada a uma classe; cada percurso na árvore (da raíz até à folha) corresponde a uma regra de classificação. No espaço definido pelos atributos cada folha corresponde a uma região (hiper-rectângulo); a intersecção dos hiper-rectângulos é vazio e a união é o espaço completo [17]. As árvores do tipo `CART` permitem uma aproximação seccionalmente constante da função objectivo, onde cada região de valor constante é delimitada por hiper-rectângulos perpendiculares aos eixos (um função "histogram-like") [18]. Cada nodo interno da árvore corresponde a uma separação das variáveis (por exemplo a característica de um portátil), sendo o valor da separação seleccionado o que minimiza o erro de adaptação ("error of fit"). A figura 2.2 corresponde a uma das árvores obtidas num caso de estudo onde são consideradas as características dos filmes para realizar recomendações, no caso concreto trata-se da característica referente ao director do filme (DIR - representa a classificação média dada por um utilizador aos filmes de determinado director); o número nos círculos ou rectângulos corresponde à classificação média dos filmes que caiem no nodo [18].

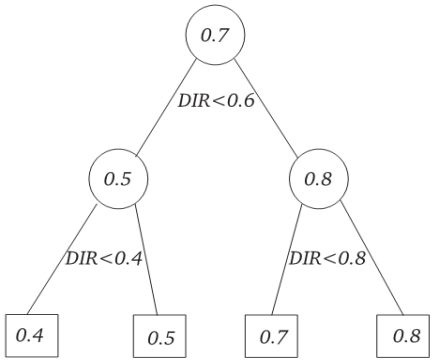


Figura 2.2: Árvore `CART` caso de estudo de filmes [18].

A utilização das probabilidades *Bayesian* foi realizada no contexto da recomendação de livros, em que cada livro é descrito por vectores de informação, sendo cada posição do vector caracterizada pelo tipo de informação (titulo, autores, sinopses, critica literária, comentários dos utilizadores, autores relacionados, títulos relacionados e assunto). Por sua vez cada tipo é tratado como um conjunto de palavras. Inicialmente é pedido aos utilizadores para pesquisarem e classificarem determinados autores e títulos de livros numa escala de 1 a 10. Nesta abordagem assume-se que a probabilidade do aparecimento de uma palavra é dependente da classe do livro e independente da frase e posição em que se insere a palavra. Também se assume que a classe c é negativa (1-5) ou positiva (6-10). Considerando a classe c_j, a palavra pertencente ao vocabulário ($a \in V$), o conjunto de palavras existente na posição m do vector (s_m), a probabilidade de cada palavra i do conjunto de palavras na posição m do vector, dado a classe e a posição $P(a_{mi}|c_j, s_m)$, a equação que estima a probabilidade da classe "à posteriori" para o livro B é [19]:

$$P(c_j|B) = \frac{P(c_j)}{P(B)} \prod_{m=1}^{S} \prod_{i=1}^{|d_m|} P(a_{mi}|c_j, s_m) \tag{2.11}$$

O S corresponde ao tamanho do vector e $|d_m|$ ao tamanho do conjunto de palavras. O "ranking" recomendado é obtido a partir da ordenação da "strength" (força) que mede o quanto é mais provável que uma palavra numa posição do vector, apareça num livro positivamente classificado do que num livro negativamente classificado. A "strength" é dada pela equação [19]:

$$strength(a_i, s_m) = \log\left(P(a_i|c_1, s_m)/P(a_i|c_0, s_m)\right) \tag{2.12}$$

A figura 2.3 mostra à direita um exemplo do "ranking" inerente à recomendação do livro "The Science of Parallel Universes - And Its Implications" e respectivo valor da "strength". No lado esquerdo é apresentado como foi obtida a "strength" associada à palavra "universes", onde surge o número de vezes que a palavra surge no vector que caracteriza cada um livros cujo título também é apresentado. Pode-se ver ainda a classificação dada pelo utilizador nesses livros [19].

Slot	Word	Strength
WORDS	MULTIVERSE	75.12
WORDS	UNIVERSES	25.08
WORDS	REALITY	22.96
WORDS	UNIVERSE	15.55
WORDS	QUANTUM	14.54
WORDS	INTELLECT	13.86

Title	Rating	Count
The Life of the Cosmos	10	15
Before the Beginning : Our Universe and Others	8	7
Unveiling the Edge of Time	10	3
Black Holes : A Traveler's Guide	9	3
The Inflationary Universe	9	2

Figura 2.3: Exemplo de recomendação [19].

Limitações

Esta abordagem baseada no conteúdo tem como limitações:

- Obtenção das características dos itens;
- Super-especialização;
- Problema do novo utilizador.

A primeira desvantagem está relacionada com a facilidade/dificuldade com que se obtêm as características dos itens, isto é, se as características se podem obter automáticamente (caso estejam em texto) ou obtidas manualmente. As técnicas de "Information retrieval" aplicam-se facilmente a características em texto, mas não noutros domínios como multimédia (por exemplo imagens, ficheiros áudio e vídeo) pois as características têm que ser introduzidas manualmente, o que não é prático e consome muitos recursos. Outra questão relacionada com esta limitação prende-se com o facto de a obtenção automática das características poder seleccionar as mesmas características para dois produtos diferentes, tornando-os indistinguíveis. Esta situação pode ser originada pelo facto de os itens serem identificados pelas palavras-chave mais importantes, que no caso de coincidirem, tornam os itens indistinguíveis [12].

A segunda limitação está relacionada com a capacidade de o `Sistemas de Recomendação` gerar recomendações considerando apenas os itens com melhor avaliação (tendo em conta o utilizador), o que conduz à recomendação de itens bastante semelhantes aos já avaliados. Por exemplo, recomendar todos os filmes de "Woody Allen" a um utilizador que apenas gosta de um filme [12].

O problema do novo utilizador prende-se com o número insuficiente de avaliações realizadas por determinado utilizador, impedindo o SR de detectar as suas preferências, o que conduz à geração de recomendações pouco fidedignas.

2.3.2 Abordagem híbrida

Esta abordagem tem como finalidade combinar as abordagens baseada no conteúdo e *Collaborative Filtering*, com o propósito de se complementarem e suplantarem as limitações referidas anteriormente relativas aos dois sistemas. Em seguida são apresentadas as diferentes formas de combinar os dois `Sistemas de Recomendação` [12]:

1. Implementar os dois sistemas independentemente e combinar as suas recomendações;
2. Incorporar algumas características dos sistemas baseados em conteúdo, nos sistemas de *Collaborative Filtering*;
3. Incorporar algumas características dos sistemas *Collaborative Filtering*, nos sistemas baseados em conteúdo;
4. Construir uma abordagem unificada que incorpora características dos dois sistemas.

Combinar separadamente os dois sistemas

Para implementação desta técnica surgem dois cenários [12]:

- Combinar as recomendações obtidas numa única recomendação, usando uma combinação linear ou um esquema de votos.
- Avaliar para cada recomendação qual dos sistemas gera recomendações mais exactas, determinando o grau de confiança de cada uma e escolher o maior ou o mais consistente com as avaliações realizadas pelo utilizador no passado.

Adicionar características do sistema baseado no conteúdo no *Collaborative Filtering*

Neste tipo de `Sistemas de Recomendação` utiliza-se a abordagem *Collaborative Filtering* mas introduzem-se características dos sistemas baseados no conteúdo, como a manutenção do perfil relativo às características dos itens que cada utilizador já avaliou. Esta estratégia permite reduzir o problema da escassez de itens comuns e também resolve o problema dos novos itens pois podem ser sugeridos com base no perfil do utilizador. Em algumas implementações são utilizados mecanismos automáticos (*filterbots*) que realizam análises baseadas no conteúdo, funcionando assim como participantes no sistema de *Collaborative Filtering*. Perante isto, os utilizadores que tiverem preferências semelhantes aos *filterbots* obtêm recomendações mais exactas [12]. A diferença entre esta abordagem

e a da secção anterior prende-se com o facto de esta incorporar algumas características do sistema baseado no conteúdo no *Collaborative Filtering*, enquanto que na anterior os dois sistemas são utilizados separadamente escolhendo a melhor recomendação.

Adicionar características do *Collaborative Filtering* no sistema baseado no conteúdo

Os `Sistemas de Recomendação` baseados nesta abordagem utilizam tipicamente a técnica de redução da dimensionalidade aplicada ao grupo de perfis baseados no conteúdo. Utilizam *latent semantic indexing (LSI)* para criar perfis de utilizadores tipo *Collaborative Filtering*, onde os utilizadores são representados como termos de vectores, aumentando a performance comparativamente aos sistemas baseados no conteúdo [12]. A técnica `LSI` é oriunda do campo da "Information retrieval" e permite descobrir padrões relativos à forma como as palavras são utilizadas num conjunto de documentos, ao agrupar as palavras podem surgir co-ocorrências que caracterizam grupos de documentos [20]. Neste caso concreto, a técnica `LSI` será aplicada sobre o conjunto de perfis (quer sobre o perfil de *Collaborative Filtering*, quer sobre o perfil do sistema baseado no conteúdo) em vez do conjunto de documentos, com o intuito de obter as semelhanças entre perfis [20].

Desenvolver um `Sistemas de Recomendação` unificado

Esta abordagem é caracterizada pela existência de várias técnicas que são apresentadas nos parágrafos seguintes.

Uma das técnicas é caracterizada pela unificação de sistema baseado no conteúdo e de um classificador baseado em regras (*ripper*). Este é caracterizado pela capacidade de obter regras a partir de dados constituídos por conjuntos de valores (filipe $\in$ portáteis, monitores, impressoras), ou seja, conjuntos de valores que em linguagem natural querem dizer que o "filipe gosta de portáteis, monitores e impressoras" ou "a maria e o filipe gostam de portáteis" [21].

Outra das técnicas propostas mistura dos algoritmos *Bayesian* e regressões que empregam a sequência *Markov* e métodos de *Monte Carlo* para estimar parâmetros e previsões. Considera-se a informação do perfil dos utilizadores e dos itens num único modelo estatístico que estima "ratings" desconhecidos r_{ij} para o utilizador i e item j [12]:

$$\begin{aligned} &r_{ij} = x_{ij}\mu + z_i\gamma_j + w_i\lambda_i + e_{ij}, \\ &e_{ij} \sim N(0, \sigma^2), \\ &\lambda_i \sim N(0, \Lambda), \\ &\gamma_j \sim N(0, \Gamma) \end{aligned} \tag{2.13}$$

Onde e_{ij}, λ_i e γ_j são variáveis aleatórias tendo em conta o efeito do ruído, fontes de utilizadores heterogéneos não observadas e itens heterogéneos, respectivamente. x_{ij} é a matriz que contém as características dos utilizadores e itens, z_i é o vecrtor das características do utilizador e w_i o vector das características dos itens. Os parâmetros desconhecidos μ, σ^2, Λ e Γ são estimados a partir das avaliações existentes utilizando a sequência *Markov* e métodos de *Monte Carlo*. Em suma, são consideradas características dos utilizadores constituintes do perfil do utilizador, as características dos itens constituintes do perfil dos itens e a sua interacção (x_{ij}) para estimar a classificação de um item [12].

Uma outra técnica prende-se com a unificação de `Sistemas de Recomendação` baseados no conhecimento e *Collaborative Filtering*. Os SR baseados no conhecimento visam a recomendação

de itens tendo em conta algumas regras do negócio e temáticas, como por exemplo a compra de um portátil originar tipicamente a compra de uma mala e rato USB. A utilização desta técnica pretende colmatar os problemas do novo utilizador e novo item que se verificam nos *Collaborative Filtering* na fase inicial, quando ainda não existe informação suficiente para gerar recomendações fidedignas [12].

2.4 Avaliação

A avaliação visa fornecer medidas que permitem saber, num cenário empírico, o quanto as recomendações geradas pelo sistema estão de acordo com os itens que na realidade o utilizador visualizou. Pode-se assim ficar com uma ideia do comportamento do sistema quando implementado num cenário. Através da avaliação podem-se comparar diferentes abordagens e técnicas.

Das medidas de avaliação mais populares temos a *Precision* e *Recall*, surgiram em 1968 aquando dos primeiros sistemas de "Information retrieval". Estas são apresentadas através de uma tabela 2.2. O conjunto de itens deve ser separado em duas classes: itens relevantes e não relevantes.

	Seleccionados	Não Seleccionados	Total
Relevantes	N_{rs}	N_{rn}	N_r
Irrelevantes	N_{is}	N_{in}	N_i
Total	N_s	N_n	N

Tabela 2.2: Classificação dos itens

A *Precision* pode ser caracterizada como sendo a divisão do número de itens relevantes seleccionados em determinada pesquisa, pelo número de itens seleccionados pela pesquisa [23].

$$Precision = \frac{|\{itens \quad relevantes\} \cap \{itens \quad seleccionados\}|}{|\{itens \quad seleccionados\}|}$$

O *Recall* corresponde à divisão do número de itens relevantes seleccionados pela pesquisa, pelo número de itens relevantes existentes [23].

$$Recall = \frac{|\{itens \quad relevantes\} \cap \{itens \quad seleccionados\}|}{|\{itens \quad relevantes\}|}$$

Quando a *Precision* assume o valor 1 significa que todos os itens seleccionados pela pesquisa são relevantes (mas não diz nada acerca dos itens relevantes que não foram seleccionados). Quando o *Recall* apresenta o valor 1 significa que todos os itens relevantes foram seleccionados pela pesquisa (mas não diz nada acerca de quantos itens irrelevantes foram seleccionados) [23].

Devido ao facto destas duas medidas isoladas não avaliarem exaustivamente o problema, aparecem tipicamente combinadas numa medida única que é *F*, que pode ser caracterizada como sendo a média harmónica pesada da *Precision* e *Recall*. É a medida da eficácia da selecção, relativamente ao utilizador que atribui β vezes importância ao *Recall* relativamente à *Precision* [23].

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{\beta^2 * precision + recall}$$

No caso concreto do problema abordado nesta dissertação o $\beta = 1$, portanto a equação assume a seguinte forma:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

2.4.1 Metodologia de avaliação

As metodologias de avaliação têm como finalidade a criação de contextos controlados nos quais se consiga simular as condições reais, com o propósito de avaliar as metodologias implementadas.

Uma das metodologias passiveis de utilização caracteriza-se pela divisão do conjunto de dados em dois: treino e teste, com diferentes finalidades. O conjunto de treino é a parte dos dados que constitui o histórico de utilizadores a partir do qual se vão realizar as recomendações aos utilizadores activos. O conjunto de teste visa simular os utilizadores activos através da divisão das sessões em itens conhecidos (já visitados pelo utilizador) e não conhecidos, com o objectivo de os prever utilizando o conjunto de treino. Os itens previstos são depois comparados com os itens dados como não conhecidos e é feita a avaliação. De cada sessão do conjunto de teste são dados como conhecidos l itens, os restantes itens são dados como não conhecidos. Para evitar que existam sessões sem itens não conhecidos são consideradas sessões com pelo menos $l + 2$ itens, garantindo que há pelo menos dois itens não conhecidos. Para além desta metodologia, também é utilizado o parâmetros experimental "All but 1" caracterizado pela utilização de todos os itens vistos na sessão, excepto um. Considera-se assim que os itens das sessões, caracterizados como não conhecidos, representam itens relevantes. A tabela 2.3 visa exemplificar graficamente o que foi dito anteriormente, considerando que a área a cinza corresponde aos itens dados como não conhecidos. Na linha i_m é exemplificado um caso correspondente ao parâmetro "All but 1".

		j_1	..	j_l	..	j_{l+2}	j_k
Conj. treino	i_1						
	i_2	$v_{2,1}$					
	:						
Conj. teste	i_a						
	:						
	i_m						

Tabela 2.3: Esquema da organização conceptual do conjunto de treino e teste.

Os parâmetros experimentais onde são conhecidos l itens permitem obter a eficácia para sessões com menos itens conhecidos e conjunto de dados de memória de menor dimensão. Enquanto o parâmetro "All but 1" possibilita a avaliação da eficácia para sessões onde são conhecidos muitos itens e para cenários em que o conjunto de dados histórico é de maior dimensão [6].

2.5 Exemplos de Sistemas de Recomendação

A tabela 2.4 apresenta um resumo de vários estudos e técnicas utilizadas para proceder à criação de Sistemas de Recomendação ao longo dos tempos. Para cada técnica de recomendação são apresentadas duas abordagens possíveis, isto é, *Memory-Based* ou no *Model-Based*, como já foi mencionado numa das secções anterior.

Abordagens de Recomendação	**Técnicas Aplicadas**	
	Memory-Based	*Model-Based*
Baseada em Conteúdo	Técnicas mais utilizadas: ⇒ Adaptação de técnicas oriundas da área de "Information retrieval" ⇒ Agrupamento Exemplos de Pesquisas: ⇒ Lang, 1995 ⇒ Balabanovic & Shoham 1997 ⇒ Pazzani & Billsus 1997	Técnicas mais utilizadas: ⇒ *Bayesian classifiers* ⇒ *Cluster models* ⇒ Árvores de Decisão ⇒ Redes Neutrais Exemplos de Pesquisas: ⇒ Pazzani & Billsus 1997 ⇒ Billsus & Pazzani 1999, 2000 ⇒ Mooney et al. 1998 ⇒ Mooney & Roy 1999 ⇒ Zhang et al. 2002
Collaborative Filtering	Técnicas mais utilizadas: ⇒ Vizinhança mais próxima (co-seno, correlação) ⇒ Teoria dos Gráficos Exemplos de Pesquisas: ⇒ Resnick et al. 1994 ⇒ Hill et al. 1995 ⇒ Shardanand & Maes 1995 ⇒ Breese et al. 1998 ⇒ Nakamura & Abe 1998 ⇒ Aggarwal et al. 1999 ⇒ Delgado & Ishii 1999 ⇒ Pennock & Horwitz 1999 ⇒ Sarwar et al. 2001	Técnicas mais utilizadas: ⇒ *Bayesian networks* ⇒ *Cluster models* ⇒ Redes Neuronais ⇒ Regressão Linear ⇒ Modelos Probabilísticos Exemplos de Pesquisas: ⇒ Billsus & Pazzani 1998 ⇒ Breese et al. 1998 ⇒ Ungar & Foster 1998 ⇒ Chien & George 1999 ⇒ Getoor & Sahami 1999 ⇒ Pennock & Horwitz 1999 ⇒ Goldberg et al. 2001 ⇒ Kumar et al. 2001 ⇒ Pavlov & Pennock 2002 ⇒ Shani et al. 2002 ⇒ Yu et al. 2002, 2004 ⇒ Hofmann 2003, 2004 ⇒ Marlin 2003 ⇒ Si & Jin 2003
Híbrida	Combinar componentes baseados em conteúdo e *Collaborative Filtering*: ⇒ Combinação linear de avaliações previstas ⇒ Esquemas variados de votação ⇒ Incorporar um componente como parte da abordagem do outro Exemplos de Pesquisas: ⇒ Balabanovic & Shoham 1997 ⇒ Claypool et al. 1999 ⇒ Good et al. 1999 ⇒ Pazzani 1999 ⇒ Billsus & Pazzani 2000 ⇒ Tran & Cohen 2000 ⇒ Melville et al. 2002	Combinar componentes baseados em conteúdo e *Collaborative Filtering*: ⇒ Incorporar componente como parte de um modelo no outro ⇒ Construir um modelo unificado Exemplos de Pesquisas: ⇒ Basu et al. 1998 ⇒ Condliff et al. 1999 ⇒ Soboroff & Nicholas 1999 ⇒ Ansari et al. 2000 ⇒ Popescul et al. 2001 ⇒ Schein et al. 2002

Tabela 2.4: Classificação de Sistemas de Recomendação [12]

2.5.1 Projectos Académicos

GroupLens

O sistema *GroupLens* teve por base o sistema *Tapestry* que foi um dos primeiros sistemas de recomendação, cujo objectivo era ajudar pequenos grupos de trabalho, a resolver o problema de excesso de informação. Este sistema permitia a avaliação da informação por parte dos utilizadores e possibilitava também aos utilizadores especificar quais os utilizadores predilectos [8], com o propósito de terem acesso à informação acedida pelo utilizador predilecto.

O sistema *GroupLens* difere essencialmente em dois pontos: a previsão é baseada na agregação de pontuações inseridas pelos utilizadores e retirou-se a possibilidade do utilizador conhecer antecipadamente quem realizou as avaliações.

Posto isto, o sistema *GroupLens* permite ajudar grupos de trabalho a encontrar informação interessante. O sistema selecciona utilizadores cujas preferências são similares às do utilizador activo.

O diagrama apresentado na figura 2.4 mostra a arquitectura do sistema, constituído pelo *GroupLens Ratings Bureau* que funciona como um *broker* de pedidos, para proporcionar uma utilização distribuída do sistema. Neste diagrama há dois clientes: o *xrn* já estabeleceu ligação e já lhe foi atribuído o processo de previsão e o cliente *tin* que também já estabeleceu ligação mas ainda não lhe foi atribuído processo de previsão.

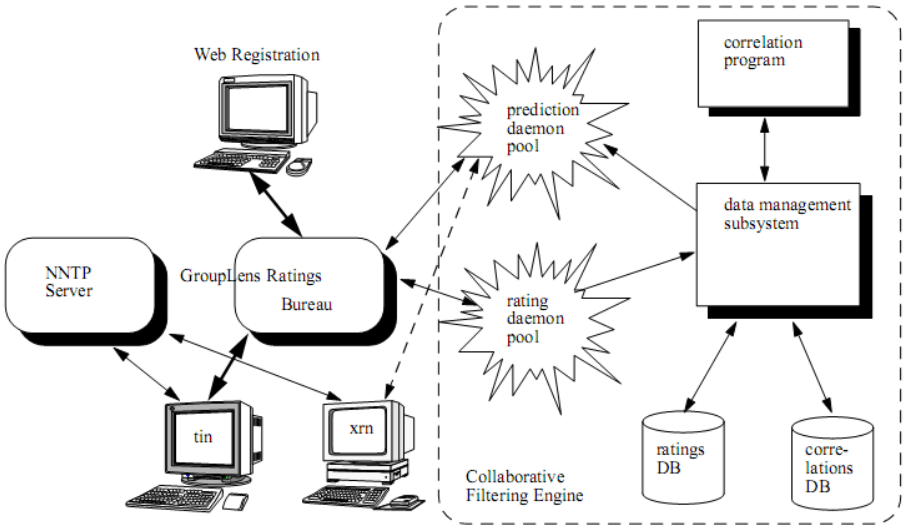


Figura 2.4: Diagrama da arquitectura do *GroupLens*. [8]

O módulo de previsão é constituído por quatro sub-módulos:

- fila de processos de previsão (*prediction deamon pool*) ⇒ trata os pedidos para obtenção de previsões;
- fila de processos de avaliação (*rating deamon pool*) ⇒ possibilita a recepção das avaliações realizadas pelos utilizadores;

- programa de correlação (*correlation program*) ⇒ determina a semelhança entre os utilizadores, através da informação avaliada no passado por estes (*Collaborative Filtering*);
- subsistema de gestão de dados (*data management subsystem*) ⇒ é responsável pela gestão das base de dados das avaliações e correlações.

RINGO

O sistema *RINGO* foi desenvolvido para recomendação personalizada de música. Este trabalho explora semelhanças entre as características dos itens considerados nas preferências dos diferentes utilizadores com o intuito de recomendar itens. Isto baseia-se no facto de as preferências das pessoas apresentarem tendências gerais e padrões entre os grupos de indivíduos [14].

Neste sistema os utilizadores descrevem as suas preferências musicais, através da avaliação de algumas músicas. Estas avaliações constituem o perfil dos utilizadores, sendo estes perfis considerados na geração das recomendações para cada utilizador [14].

O sistema *Ringo* comparava os perfis dos utilizadores para determinar quais os que apresentavam gostos similares (gostavam dos mesmos álbuns e/ou não gostavam dos mesmos, por exemplo). Inicialmente os utilizadores similares são identificados a partir da comparação de perfis. O sistema pode predizer o quanto o utilizador gosta de um álbum/artista que ainda não foi avaliado pelo mesmo [14]. Para determinar o perfil de cada utilizador, quando este acede pela primeira vez ao *Ringo* é-lhe apresentada uma lista de 125 artistas para os avaliar de acordo com o quanto gosta de os ouvir. Caso o utilizador não esteja familiarizado com o artista ou não possuir uma opinião formada sobre o mesmo, este é incitado a não avaliar, para não ocorrerem distorções do perfil. A pontuação segue uma escala de avaliação de 7 pontos, sendo o valor 1 (não gosta), 4 (indiferente) e 7 (adora) [14].

A lista de artistas mencionada anteriormente está dividida em duas partes. Uma parte da lista é constituída pelos artistas mais pontuados, isto assegura que um novo utilizador tem a oportunidade de pontuar artistas que outros já pontuaram, de forma a existirem pontuações em comum entre os perfis dos utilizadores. A outra parte da lista é obtida através de uma selecção aleatória. Após submissão das avaliações que constituem o perfil inicial do utilizador, o *Ringo* pode realizar a geração de previsões, ou seja, um utilizador pode pedir ao *Ringo* para [14]:

1. sugerir novos artistas/álbuns que o utilizador gostaria de obter ou ouvir;
2. listar artistas/álbuns que o utilizador não gostaria;
3. realizar uma previsão sobre um artista/álbum específico.

O retorno dado pelo *Ringo* aos utilizadores não incluí nenhuma informação em particular sobre a identidade dos outros utilizadores que contribuíram para as recomendações. É de destacar que em sistemas *Collaborative Filtering* a identidade de quem avaliou deve ser mantida em segredo.

O sistema *Ringo* também permite a escrita de comentários sobre o produto recomendado. Os próprios utilizadores têm a opção de inserir novos artistas e álbuns e receber mensagens sobre os novos itens e novidades do sistema.

Ao contrário do sistema *GroupLens*, o *Ringo* segue uma abordagem baseada no conteúdo.

MovieLens

O sistema *MovieLens* foi criado para realizar recomendação de filmes, onde o utilizador tem que efectuar algumas avaliações de filmes, de modo a formar o seu perfil das preferências e encontrar os utilizadores com preferências semelhantes [25]. Embora neste ponto se verifique semelhança com no sistema apresentado anteriormente (*Ringo*), a abordagem utilizada por este sistema para gerar as recomendações é *Collaborative Filtering*, tal como o sistema *GroupLens*, onde as previsões são realizadas com base na semelhança dos perfis dos utilizadores e não com base na semelhança das características dos itens que constituem os perfis [28].

O utilizador pode pedir previsões para determinado filme, para tal tem que seleccionar um filme que ainda não avaliou e pedir que seja gerada uma previsão com base no seu perfil. Na figura 2.5 é apresentado um exemplo de recomendações através das estrelas preenchidas a vermelho. O utilizador também pode proceder à sua avaliação, através da caixa de selecção [25].

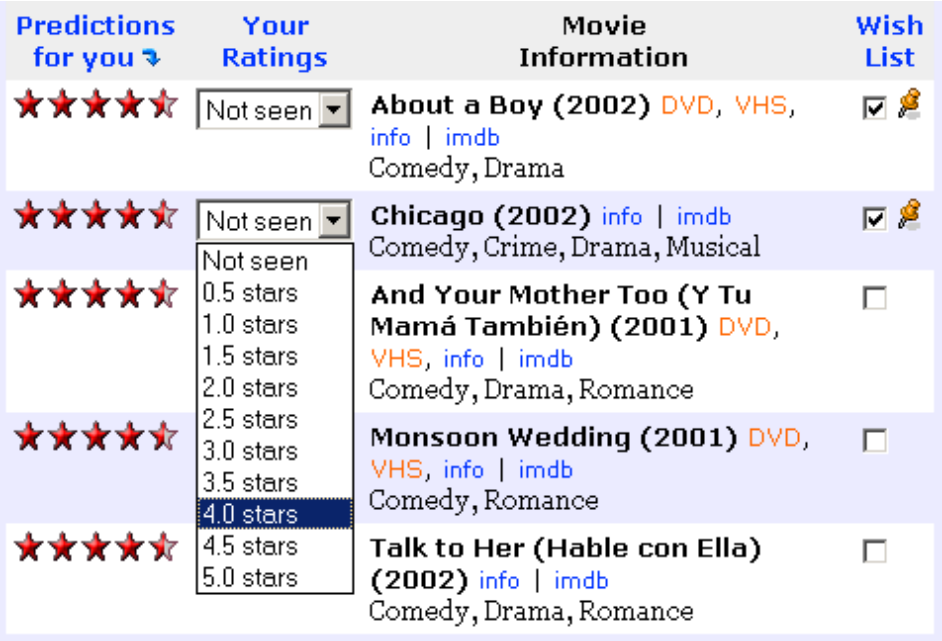


Figura 2.5: Exemplo de recomendação [25]

Uma funcionalidade muito interessante é a possibilidade de criar grupos de utilizadores de forma a obter recomendações para esse grupo [25].

2.5.2 Sítios Comerciais

Amazon.com

Este é um dos sítios mundialmente conhecidos que utiliza `Sistemas de Recomendação` e que permite a venda *on-line* de diversas categorias de produtos. Neste sítio podem-se encontrar os seguintes tipos de recomendações [24]:

- Clientes que viram ⇒ nesta área são apresentados os itens que foram vistos por outros uti-

lizadores que também viram o item que o utilizador activo está actualmente a observar.

- Clientes que compraram ⇒ esta área caracteriza-se por disponibilizar os itens que foram adquiridos por utilizadores que também compraram o item que o utilizador activo está a visualizar.
- Classificação dos itens ⇒ Possibilita a classificação dos itens, conhecer a sua classificação média e discriminar as diversas classificações.
- Comentários sobre os itens ⇒ Permite ao utilizador obter a opinião de outros utilizadores, aconselhando ou desaconselhando a compra dos itens.

Os dois primeiros pontos referem-se a recomendações personalizadas porque têm em conta o que o utilizador está a visualizar em determinado momento. No primeiro ponto as recomendações são feitas a partir do histórico de compras e avaliações feitas pelo utilizador activo e a recomendação é obtida através de *Collaborative Filtering* [16]. No segundo ponto a recomendação é realizada a partir dados da implícitos acerca dos itens visualizados pelo utilizador e a recomendação também é obtida através de *Collaborative Filtering* [16].

Os dois últimos tipos de recomendação não são personalizados, uma vez que são comuns a todos os utilizadores. As recomendações são obtidas a partir do tratamento estatístico dos dados relativos a todos os utilizadores, como histórico de compras, avaliações e comentários [16].

eBay

É um sitio que utiliza `Sistemas de Recomendação` e possibilita a realização de leilões *on-line*. Permite assim o contacto directo entre os vendedores e os compradores, cabendo ao *eBay* a gestão de todo o processo da transacção.

Devido à imensa variedade de itens que estão à venda, este sitio tornou-se um óptimo local para compras *on-line*, permitindo satisfazer grande parte das necessidades dos utilizadores. Para além da grande diversidade de itens também são apresentadas informações semelhantes às lojas *on-line*, como fotografias e descrição detalhada. É ainda disponibilizada informação relativa ao preço nas lojas e o preço de rua, possibilitando assim uma melhor avaliação da mais-valia da compra.

Este sitio tem sido também utilizado por coleccionadores, para venda e compra de itens raros e antigos. Outro ponto interessante é ser completamente grátis para os compradores e pouco dispendioso para quem vende.

No *eBay* os vendedores podem colocar à venda quase todos os tipos de itens, desde carros a livros. O vendedor tem disponíveis três tipos de métodos de venda:

- *Auction-style listings* ⇒ possibilita a venda de um ou mais itens durante um determinado período de tempo. O vendedor pode estabelecer um preço mínimo.
- *Fixed Price format* ⇒ permite vender um ou mais itens de uma forma imediata. Isto é, a transacção é realizada no momento em que o comprador aceita o preço pedido. Neste caso não há mais nenhuma licitação do produto, basicamente é uma compra no sítio *on-line*.
- *Dutch Auctions* ⇒ este leilão é especialmente desenhado para a venda de vários itens semelhantes no mesmo leilão. As propostas podem ser realizadas para 1 ou mais itens incluídos no leilão.

Devido ao elevado número de utilizadores e quantidade de informação, é necessário garantir acesso rápido aos itens, para tal são disponibilizadas as seguintes funcionalidades no sítio: os vizinhos relacionados, os guias relacionados e os itens recentemente adicionados às categorias relacionadas mais vistas.

A primeira funcionalidade tem como intuito a sugestão de "Neighborhoods Blog" relacionados com o tema que o utilizador activo está a visualizar. A segunda possibilita a sugestão de guias relacionados com a temática que o utilizador activo está a consultar. A terceira funcionalidade, visa apresentar os itens adicionados recentemente, através da identificação das categorias mais vistas, relacionadas com a categoria que esta ser consultada pelo utilizador activo.

O *eBay* possibilita ainda o envio de e-mail's com informação relativa aos leilões que potencialmente podem interessar ao utilizador. Esta informação é obtida a partir das características dos itens que o utilizador já esteve interessado no passado, trata-se assim de uma abordagem baseada no conteúdo.

Drugstore

Este sítio dedica-se à venda de itens da área da saúde e cosméticos. É um sítio com um volume de vendas bastante elevado, principalmente ao nível do mercado dos Estados Unidos da América [27]. Quando o utilizador selecciona determinado item, é apresentada a avaliação atribuída pelos outros utilizadores, assim como os comentários que estes fizeram. A obtenção desta informação é feita a partir de sumarização estatística [16]. Também é disponibilizada uma lista de produtos relacionados tendo em conta o produto seleccionado. A obtenção desta lista é feita a partir de recomendação baseada no conteúdo ou selecção manual [16].

2.6 Desafios dos `Sistemas de Recomendação`

Para além de ultrapassar as limitações apresentadas nas secções anteriores, é ainda necessário ter em conta os seguintes desafios [4]:

- *Scalability* ⇒ refere-se à capacidade que a abordagem demonstra em lidar com sítios com elevada actividade e quantidade de dados, isto é, os `Sistemas de Recomendação` têm que ser eficientes de forma a minimizarem o tempo utilizado na geração de recomendações, assim como na utilização de recursos de *hardware*. Este ponto é potencialmente crítico pois as recomendações devem ser realizadas aquando da geração da página para que o utilizador as possa visualizar. A geração de recomendações não deve contribuir para a diminuição da capacidade de resposta do sítio.

- Evolução dos itens e dos interesses dos utilizadores ⇒ lidar com a rápida evolução dos interesses dos utilizadores e com a constante evolução dos itens disponíveis no sítio (quer ao nível das características e funções, quer ao nível de marcas).

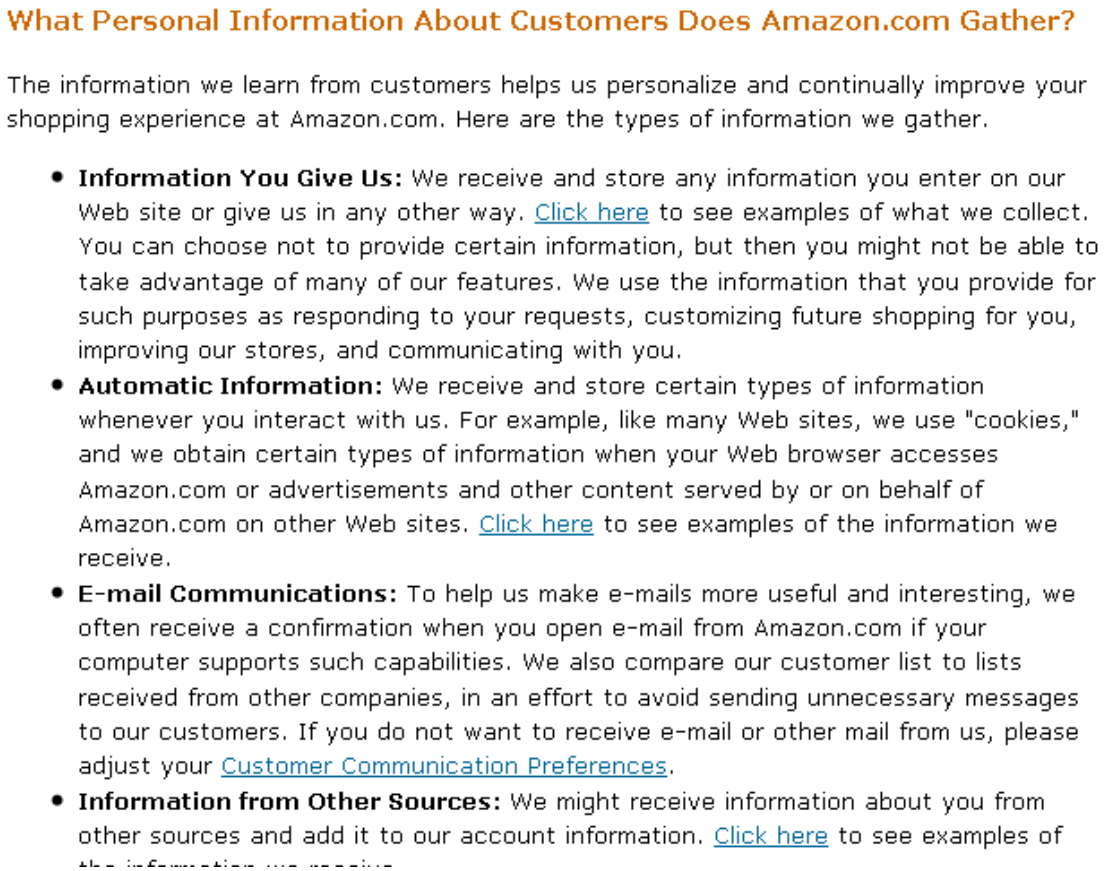
- Integração de fontes de dados múltiplas ⇒ Utilizar várias fontes de dados para além dos tradicionais *Web Access Logs*, como o conteúdo das páginas visitadas e as ligações que existem entre as várias páginas. Pretendem-se assim conhecer os padrões de navegação dos utilizadores.

2.6.1 Questões de privacidade

Para que um `Sistemas de Recomendação` possa proceder à recomendação, tem que recolher informação relativa aos utilizadores, tal como os itens que visitou/avaliou, qual a avaliação que deles fez, itens que comprou, dados demográficos, etc. Só tendo alguns destes dados é que os SR podem realizar recomendações exactas, de forma a melhorar a satisfação dos utilizadores/clientes. Algumas destas informações são recolhidas de forma implícita, ou seja, sem que o utilizador se aperceba, e muitas vezes sem que esteja consciente de que tal informação está a ser recolhida.

Para tratar destas questões a *World Wide Web Consortiun* (W3C) propôs algumas regras para servirem de guia, designada por *Platform for Privacy Preferences* (P3P), que possibilita aos sítios exporem as suas directivas de privacidade de uma forma simples e de fácil compreensão [4]. Estas directivas também podem ser reconhecidas pelo *browser* de modo a que sejam bloqueadas determinadas funções caso o sítio não implemente este guia [29].

O sítio da *Amazon* exemplifica a preocupação que este tema merece. Nele todas as questões mais sensíveis para os utilizadores são abordadas e clarificadas, para que não haja dúvidas e de forma a promover o sentimento de segurança dos dados fornecidos implícita ou explicitamente. Na figura 2.6 é apresentado um trecho da politica de privacidade. Esta informação está disponível para consulta no sítio.



What Personal Information About Customers Does Amazon.com Gather?

The information we learn from customers helps us personalize and continually improve your shopping experience at Amazon.com. Here are the types of information we gather.

- **Information You Give Us:** We receive and store any information you enter on our Web site or give us in any other way. Click here to see examples of what we collect. You can choose not to provide certain information, but then you might not be able to take advantage of many of our features. We use the information that you provide for such purposes as responding to your requests, customizing future shopping for you, improving our stores, and communicating with you.
- **Automatic Information:** We receive and store certain types of information whenever you interact with us. For example, like many Web sites, we use "cookies," and we obtain certain types of information when your Web browser accesses Amazon.com or advertisements and other content served by or on behalf of Amazon.com on other Web sites. Click here to see examples of the information we receive.
- **E-mail Communications:** To help us make e-mails more useful and interesting, we often receive a confirmation when you open e-mail from Amazon.com if your computer supports such capabilities. We also compare our customer list to lists received from other companies, in an effort to avoid sending unnecessary messages to our customers. If you do not want to receive e-mail or other mail from us, please adjust your Customer Communication Preferences.
- **Information from Other Sources:** We might receive information about you from other sources and add it to our account information. Click here to see examples of

Figura 2.6: Trecho da politica de privacidade do sítio da *Amazon* [24].

Alguns investigadores também já realizaram algumas tentativas para manter a privacidade, através do mascarar dos dados do utilizador com métodos como aleatoriedade e modificando os dados introduzidos, sem alterar significativamente os resultados [4].

Capítulo 3

Enriquecimento dos dados para *CF*

No contexto de recomendação de produtos on-line, frequentemente está disponível informação que não é usada no *Collaborative Filtering* e que é potencialmente útil para melhorar as recomendações. Por exemplo os produtos que são utilizados como itens estão frequentemente organizados em hierarquias de categorias que também são alvo de acesso por parte dos utilizadores e registadas nos *Web Access Logs*. Adicionalmente, pode-se ainda considerar o intervalo de tempo que o utilizador demora na visualização de uma página, que pode ser indicativo do seu interesse no produto. Assim, para além do estudo comparativo do algoritmo de correlação e suas variantes *Inverse User Frequency* e *Default Vote*, vão ser apresentadas formas de integrar esta informação no algoritmo *CF*.

A razão que conduziu à utilização da informação relativa às hierarquias de categorias prende-se com o facto de ser de fácil obtenção, estar disponível no caso de estudo e de ser transversal à maioria dos sítios de comercio electrónico. Por outro lado a integração desta informação também pode ter impacto na minimização do problema do novo item, visto que as categorias são menos voláteis que os produtos.

Relativamente ao intervalo de tempo de visualização de uma página relativa a um produto também é transversal e de fácil obtenção, uma vez que a referência temporal de acesso a cada página está presente no *Web Access Logs*, permitindo obter este intervalo a partir da diferença entre o acesso à página do produto e a página visitada posteriormente. Segundo [7], há uma relação entre o intervalo de tempo que um utilizador demora na visualização de determinado conteúdo e o interesse ou preferência por esse conteúdo.

Neste capítulo, começamos por descrever a aplicação que motivou esta abordagem. Depois discutimos os métodos de integração de informação propostos. Finalmente, apresentamos os resultados da análise exploratória dos dados.

3.1 A Empresa

A Suprides é uma empresa que faz importação, *assemblagem* e comercialização de material informático, prestando também serviços de assistência técnica.

A sua actividade começou em 1991 com a denominação 'Oporto BBS' e tem vindo a usar todo o conhecimento reunido ao longo deste tempo para apostar na procura das soluções mais adequadas a cada um dos seus clientes, porque a sua satisfação é um dos seus principais objectivos. Durante

este percurso conquistou uma vasta carteira de clientes distribuídos pelos mais diversos sectores de actividade.

As tecnologias de informação são uma peça fundamental para imprimir uma nova dinâmica nas empresas, independente do seu sector de actividade e, como tal, a Suprides tem um conjunto completo de soluções (voz, dados, vídeo, `Internet`, segurança, consultoria e *outsourcing*).

As soluções disponibilizadas estendem-se desde as redes privadas virtuais, *VPN IP*, passando por serviços de Voz sobre IP, a instalação e gestão de servidores de voz e dados, vídeo-conferência, redes geridas incluindo as *LAN* dos clientes, serviços de mobilidade *Wi-Fi* e soluções de segurança.

No que diz respeito à evolução e crescimento, esta é a primeira empresa na sua área geográfica a dinamizar os seus produtos e negócios numa óptica de divulgação, venda e assistência técnica, realizando acções de consultadoria (telefónica ou pessoal) na pré e pós-venda (Ensino - Faculdades e Feiras - Exponor).

É desde 1994 uma das empresas que mais vende computadores, impressoras e peças, no norte do país e ponto de referência a nível nacional, tendo sido já premiada pelas seguintes revistas:

- Bit;
- Exame Informática;
- PcGuia;
- Personal Computer World;
- Pro-Teste

A Suprides tem como um dos seus principais objectivos a satisfação dos clientes, impondo-se altos padrões de qualidade quer na prestação de serviços, quer na venda de material informático.

Outro objectivo é o crescimento constante do volume de negócios, para tal recorreu a três estratégias que pretende prolongar no tempo:

- Criação de parcerias em vários pontos do país;
- Criação de um departamento orientado para Empresas;
- Desenvolvimento de uma infraestrutura `WEB` abrangente a todas as áreas de negócio.

3.2 Volume de dados

O gráfico 3.1 apresenta as estatísticas do sítio `supride21` referentes ao ano de 2006, neste pode constatar-se um maior número de visitantes no início do ano. É de salientar o considerável número de acessos anuais e os *GigaBytes* transferidos.

O gráfico 3.2 apresenta as estatísticas do sítio `supride21` referentes ao ano de 2007, também se pode verificar o maior número de visitantes no início do ano.

Figura 3.1: Estatísticas de 2006.

Mês	Visitantes únicos	Numero de visitas	Páginas	Hits	Bytes
Jan 2006	12293	16638	168540	1736534	8.44 GB
Fev 2006	9575	12912	124649	1423647	7.41 GB
Mar 2006	10650	14548	136204	1516896	7.81 GB
Abr 2006	8823	11659	101886	1112726	5.91 GB
Mai 2006	8448	11184	104400	990550	5.55 GB
Jun 2006	7913	10491	94602	927956	5.31 GB
Jul 2006	7210	9782	97850	945931	4.61 GB
Ago 2006	8660	11584	100063	1009347	4.78 GB
Set 2006	9525	12741	120303	1241726	5.90 GB
Out 2006	8307	11266	114562	1046823	5.12 GB
Nov 2006	6439	8723	81701	787893	4.29 GB
Dez 2006	7229	9445	83230	888096	4.76 GB
Total	105072	140973	1327990	13628125	69.87 GB

Figura 3.2: Estatísticas de 2007.

Mês	Visitantes únicos	Numero de visitas	Páginas	Hits	Bytes
Jan 2007	7969	10331	84412	832194	4.45 GB
Fev 2007	5866	7593	57102	580686	2.92 GB
Mar 2007	6160	7976	58127	623113	2.93 GB
Abr 2007	5547	7190	53752	551557	2.49 GB
Mai 2007	4746	6362	55056	519725	2.54 GB
Jun 2007	5733	7445	56590	531976	3.10 GB
Jul 2007	6062	7521	48271	546213	2.51 GB
Ago 2007	5373	6701	49063	509308	2.14 GB
Set 2007	4005	4993	40857	444140	1.96 GB
Out 2007	0	0	0	0	0
Nov 2007	0	0	0	0	0
Dez 2007	0	0	0	0	0
Total	51461	66112	503230	5138912	25.03 GB

3.3 Informação disponível

A informação disponibilizada para realização deste projecto consiste em *Web Access Logs* do servidor do sítio `WEB`, produtos, subcategorias e categorias, lista de produtos desejados, encomendas, registo e *login* dos utilizadores. A lista de produtos desejados possibilita aos utilizadores criar listas de produtos desejados para compras futuras. Relativamente às encomendas é obrigatório o *login*, mas os itens que constituem o carro de compras são guardados independentemente do *login*.

Para a execução desta dissertação foi utilizada informação relativa aos *Web Access Logs* e aos produtos, subcategorias e categorias. Cada produto pertence a uma subcategoria e a uma categoria. Por uma questão de tempo a restante informação disponibilizada não foi utilizada.

Os *Web Access Logs* são ficheiros onde são registados todos os pedidos de dados realizados ao servidor. Cada página é constituída por vários ficheiros e todos estes ficheiros são fornecidos pelo servidor. Quando ocorre o pedido destes ficheiros ocorre a inserção de uma nova linha nos *Web*

Access Logs. A título de exemplo é apresentado em seguida um registo deste ficheiro.

```
89-180-166-23.net.novis.pt - - [31/May/2007:18:47:31 +0100]
"GET /produto.php?pid=6502& HTTP/1.1" 200 30353
 "http://www.lojasuprides.com/pesquisa.php?pag=1&ordenar=alf_az&lim=5&mid=LG&"
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Avant Browser)"
```

Nestes *logs* é registada a seguinte informação [22]:

- endereço IP (*Internet Protocol*) que pediu a informação (89-180-166-23.net.novis.pt);
- utilizador (neste caso é nulo);
- data e hora (31/May/2007:18:47:31 +0100);
- método (GET, POST);
- ficheiro requerido (/produto.php?pid=6502). Neste caso está a ser pedida uma página, mas também podem ser pedidos outros ficheiros como imagens, elementos de estilo da página (CSS - *Cascading Style Sheets*), elementos multimédia (*flash objects*, *media player* e outros), JavaScript, etc.;
- código do resultado do acesso (200 - corresponde a sucesso);
- os *Bytes* transferidos (30353);
- página que requereu (http://www.lojasuprides.com/pesquisa.php?pag=1&ordenar=alf_az&lim=5&mid=LG&);
- *browser* utilizado (Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Avant Browser)).

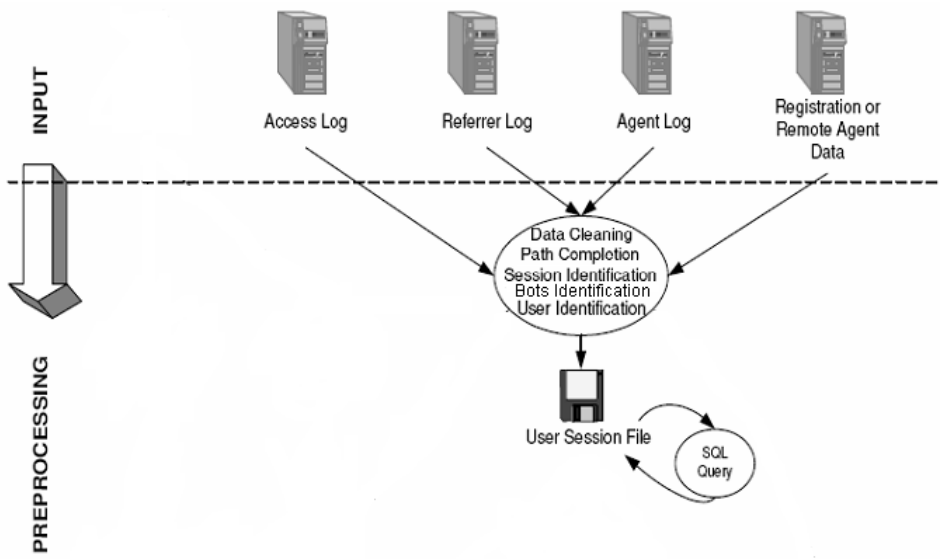
3.4 Pré-processamento dos *Web Access Logs*

Antes dos dados existentes nos *Web Access Logs* poderem ser utilizados é necessário realizar tarefas de limpeza dos dados, identificação de utilizadores e sessões, completar caminhos (*path*) e detecção de *bots*, como se pode ver na figura 3.3. Na tabela 3.1 são apresentadas algumas linhas do *Web Access Log* de forma a exemplificar o resultado de cada operação.

A limpeza dos dados consiste na eliminação das linhas dos *Web Access Logs*, que são irrelevantes para a análise que se pretende executar, neste casos os ficheiros não desejados (JavaScript, imagens, CSS, elementos multimédia, etc.), são todas as linhas excepto as que se referem aos produtos, subcategorias e categorias. Tendo em conta a tabela 3.1 esta operação levaria à exclusão das linhas 2, 3 e 4.

A identificação de utilizadores visa agregar as várias linhas presentes nos *Web Access Logs* por utilizador. Esta tarefa é executada através do endereço IP (*Internet Protocol*), embora seja passível de erro devido à utilização de ISP (*Internet Service Provider*), pois estes podem atribuir a utilizadores distintos o mesmo endereço IP. Para tentar diferenciar os utilizadores, mesmo quando esta situação se verifica, foram apresentadas duas heurísticas. Numa delas para além do IP é considerado o *browser* utilizado, isto é, assume-se que é o mesmo utilizador se o *browser* for o mesmo. Através da tabela 3.1 pode ver-se retratada a situação descrita: para o mesmo IP existem tem dois *browser* distintos (Opera e Mozilla), então temos dois utilizadores. A outra heurística pressupõe que se o utilizador aceder a uma página que não se consegue atingir (através de *link*) a partir da última página visualizada, então não

Figura 3.3: Esquema da fase de pré-processamento (baseado em [1]).



Endereço IP	Tempo	Página	Requereu	*browser*
89.155.58.32	17:36:35	sid=6478	-	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:36:35	header1.jpg	sid=6478	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:36:35	styles.css	sid=6478	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:36:35	coolmenus4.js	sid=6478	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:40:37	sid=6477	sid=6478	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:41:27	sid=6539	-	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:43:12	sid=6488	sid=6477	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:44:14	sid=6478	-	Opera/9.21 (Windows NT 5.1; U; pt)
89.155.58.32	17:46:48	sid=6482	sid=6478	Opera/9.21 (Windows NT 5.1; U; pt)
89.155.58.32	17:47:34	sid=6556	sid=6539	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	17:49:22	sid=6477	sid=6478	Opera/9.21 (Windows NT 5.1; U; pt)
89.155.58.32	18:43:02	sid=6478	-	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
89.155.58.32	18:45:02	sid=6477	sid=6478	Mozilla/4.0(MSIE 6.0;Windows NT 5.0)
livebot-65-55-208-145.live.com	18:45:23	sid=6498	-	msnbot/1.0
crawl-66-249-65-236.googlebot.com	18:50:54	sid=6482	-	Mozilla/5.0 (Googlebot/2.1;)

Tabela 3.1: Exemplo de *Web Access Logs*

é o mesmo utilizador. Tendo em conta a tabela 3.1 pode verificar-se que a sexta linha corresponde ao acesso a um artigo de categoria diferente dos acessos das linhas anteriores, pelo que pode pressupor-se que são dois utilizadores distintos. A partir das heurísticas apresentadas e da tabela 3.1 verifica-se que existem cinco utilizadores com os seguintes acessos: 6478-6477-6488-6539-6556-6478-6477, 6539-6556, 6478-6482-6477, 6498, 6482. Estas heurísticas também não estão isentas de erro, pois no caso de um utilizador aceder através de mais que um *browser* ou no caso de o utilizador aceder a determinada página através de endereço directo (isto é, sem utilizar os *links* das páginas do sítio), há a possibilidade de considerar o mesmo utilizador como utilizadores distintos [1].

Os utilizadores podem realizar várias visitas ao sítio diferidas no tempo, que consequentemente

Identificador do Item	Descrição do Item
6478	Monitor TFT 17' SAMSUNG 713BM PLUS
6477	Monitor TFT 17' SAMSUNG SYNCMASTER 732N Preto
6488	Monitor ASUS PG191 - MONITOR TFT 19
6482	Monitor 19POL. ACER 1916WAS TFT WIDE Silver
6539	Câmara Fotográfica CANON IXUS850IS
6556	Câmara Fotográfica NIKON COOLPIX 4500

Tabela 3.2: Descrição dos itens

ficam registadas nos *Web Access Logs*. A identificação de sessões visa dividir as páginas acedidas pelo utilizador em sessões individuais, permitindo a determinação de que páginas são visualizadas, em que ordem e por quanto tempo, em suma é tudo o que ocorre durante a visita de um utilizador [1]. A forma mais simples de conseguir realizar esta operação é especificar um tempo limite, entre pedidos ao sítio, caso o pedido ultrapasse o limite considera-se como sendo outra sessão. Tipicamente utiliza-se como tempo limite 30 minutos. Há outras abordagens que calculam o tempo limite a partir das estatísticas de utilização e análise dos *Web Access Logs*, com o propósito de modelarem a utilização do sítio [1]. Por exemplo: se o sítio disponibilizar artigos científicos, é natural que o tempo limite de 30 minutos não se adeqúe, pois o tempo necessário para ler o artigo pode ser superior ao limite estabelecido. Aplicando o tempo limite de 30 minutos aos acessos do utilizador com maior número de itens visualizados, referido no paragrafo anterior, permite a divisão dos acessos em duas sessões: 6478-6477-6488-6539-6556 e 6478-6477.

Completar os caminhos é uma tarefa que visa identificar as referências a páginas não registadas no *Web Access Log*, devido à "memorização" (*caching*) que ocorre no *browser* e ISP. A realização desta tarefa implica o conhecimento da estrutura do sítio. A estrutura é necessária para verificar se a partir de determinada página se consegue aceder à próxima página que surge na sessão do utilizador [1]. Por exemplo: supondo que a página relativa ao item 6477 só é acessível a partir da página do item 6478, então verifica-se que para ir da página 6482 para a página 6478 é necessário que o utilizador tenha retrocedido para a página 6478. Posto isto tem que se adicionar esta página à sessão , passando de 6478-6482-6477 para 6478-6482-6478-6477.

Por fim temos a tarefa de detecção de *bots* cujo objectivo é identificar os acessos automáticos ao sítio, de forma a excluí-los, pois o facto de não serem utilizadores humanos iria enviesar o estudo. A detecção dos *bots* nos *Web Access Logs* pode ser facilmente realizada através do nome do servidor de acesso às diferentes páginas dos itens. As palavras mais utilizadas para identificação são: "crawl" e "bot". Na tabela 3.1 são apresentados dois exemplos de *bots* livebot-65-55-208-145.live.com e crawl-66-249-65-236.googlebot.com que seriam identificados e removidos utilizando esta metodologia. No *Web Access Log* utilizado no caso de estudo detectaram-se 2129 itens acedidos em sessões pertencentes a *bots*.

Outra forma de detecção dos *bots* nos *Web Access Logs* consiste na verificação do número de itens vistos em cada sessão, pressupondo que sessões com elevado número de acessos a itens constituem muito provavelmente sessões de *bots*. Encontrou-se no *Web Access Log* do caso de estudo uma sessão com 289 produtos visitados (o que é um número bastante exagerado), como tal, não foi considerada.

Neste processo de detecção e remoção de *bots* é necessário considerar o administrador do sítio, que embora não seja um *bot* apresenta um comportamento diferente dos utilizadores que pretendem realizar compras, como tal é preciso identificar o seu utilizador, para remover as respectivas sessões. A partir da análise do *Web Access Log* verificou-se que os acessos do administrador podem ser identificados através da coluna do *Web Access Log* referente ao "utilizador". Posto isto, constatou-se a

existência de 1889 sessões. Para mais detalhe consultar o capítulo A em anexo.

3.5 Base de dados implementada

Em seguida pretende-se apresentar de forma resumida a base de dados implementada com intuito de acolher os dados que foram utilizados pelo `Sistema de Recomendação` implementado. O diagrama contido na figura 3.4 tem como intuito apresentar as tabelas e respectivas ligações.

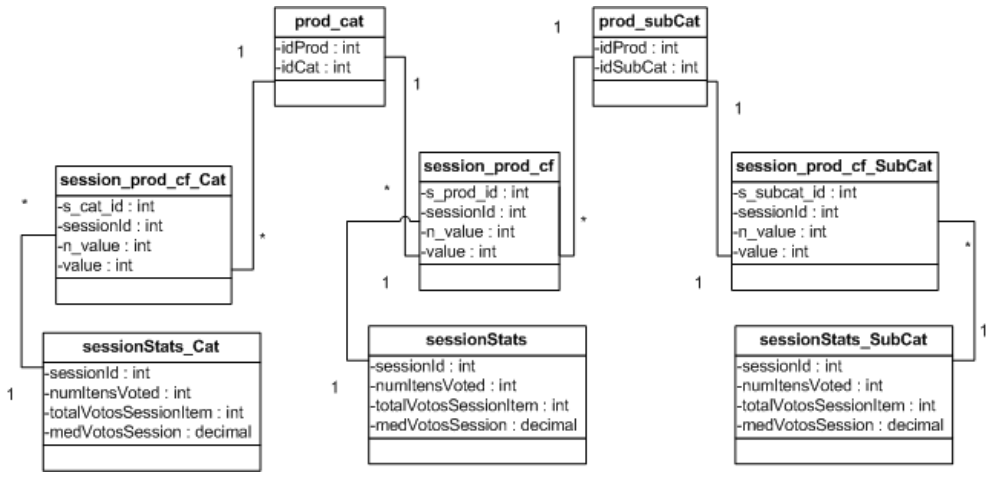


Figura 3.4: Diagrama Base de Dados Itens, Categorias e Subcategorias

A tabela "session_prod_cf_Cat" contém informação relativa aos acessos em cada sessão às categorias. É utilizada na obtenção dos utilizadores mais semelhantes ao utilizador activo tendo em com as categorias acedidas. A tabela "sessionStats_Cat" é utilizada para conter informação relativa ao calculo da média dos acessos ($\overline{v}_i$) às categorias de determinada sessão. Esta estrutura foi transposta para albergar os dados relativos aos produtos ("session_prod_cf" e "sessionStats") e subcategorias ("session_prod_cf_SubCat" e "sessionStats_SubCat"). A tabela "prod_cat" permite estabelecer a relação entre os itens e a respectiva categoria a que pertencem. A tabela "prod_subCat" é relativa às subcategorias e tem a mesma função que a tabela anterior.

Para além destas tabelas existe ainda outra que importa referir que contém os dados relativos ao tempo e é utilizada no cálculo da utilidade do tempo.

Para mais informação acerca desta temática é necessário consultar o capítulo B em anexo.

3.6 Métodos de integração de informação

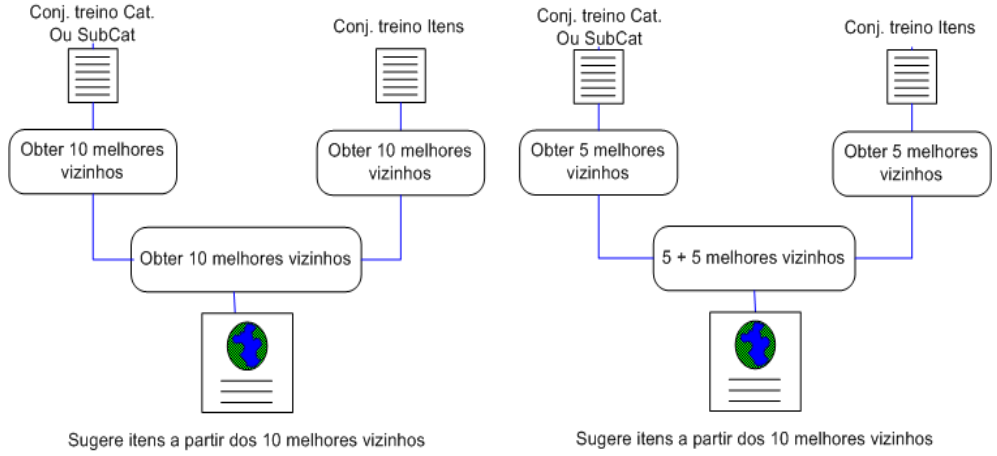
Esta secção visa apresentar quais os métodos utilizados para se proceder à integração de informação no CF relativa ao intervalo de tempo que um utilizador permaneceu na visualização de um item e a informação relativa aos acessos a categorias e subcategorias dos itens. Com integração desta informação complementar pretende-se verificar se contribui para a melhoria dos resultados, relativamente aos resultados obtidos quando utilizada apenas informação referente aos acessos a itens.

3.6.1 Integração dos acessos às Categorias e Subcategorias

Para além da informação relativa aos itens, não se pode descurar a informação relativa às subcategorias e categorias, quer ao nível da relação existente com os itens, quer ao nível dos acessos existentes nos *Web Access Logs*. Esta informação foi integrada de forma a prever os itens visitados, com o intuito de verificar se este acréscimo de informação tem influência nos resultados obtidos. Para realizar esta tarefa procedeu-se à inclusão dos dados relativos à categoria e subcategoria dos itens visitados. Cada item está agrupado por subcategoria a uma nível mais especifico e a um nível mais geral por uma categoria.

O processo adoptado para utilização da informação referida anteriormente, implica a execução de dois passos distintos, num dos quais se procede ao cálculo dos 10 utilizadores mais semelhantes ao utilizador activo (vizinhos, tendo em conta os algoritmos apresentados na secção 2.2.2) a partir do conjunto de dados dos itens, tendo em conta os itens que o utilizador activo já visitou; no outro passo procede-se há obtenção dos 10 utilizadores mais semelhantes ao utilizador activo (vizinhos) a partir do conjunto de dados contendo os acessos às categorias ou subcategorias e das categorias e subcategorias nas quais se enquadram os itens que o utilizador activo já visitou. O peso associado a cada um dos utilizadores mais semelhantes (vizinhos), obtido nos dois passos anteriores, é então comparado, de forma a escolher os melhores 10 vizinhos. Depois de escolhidos os melhores vizinhos, procede-se à obtenção dos itens a sugerir. O esquema que permite ilustrar o que foi referido anteriormente é apresentado no diagrama da figura 3.5 à esquerda.

Figura 3.5: Diagrama de integração da informação relativa às categorias ou subcategorias.



Para além do processo descrito no parágrafo anterior foi utilizado um outro caracterizado pelo facto de proceder à escolha dos 5 melhores vizinhos obtidos a partir do conjunto de dados dos itens e os 5 melhores vizinhos provenientes do conjunto de dados das categorias ou subcategorias. Constitui-se assim o conjunto de 10 vizinhos a partir dos quais se executa a sugestão dos itens.

É de notar que a correlação entre categorias ou subcategorias é distinta da correlação entre itens, isto devido há grande diferença entre o número de itens acedidos (2570) relativamente ao número de categorias (10) e subcategorias (187).

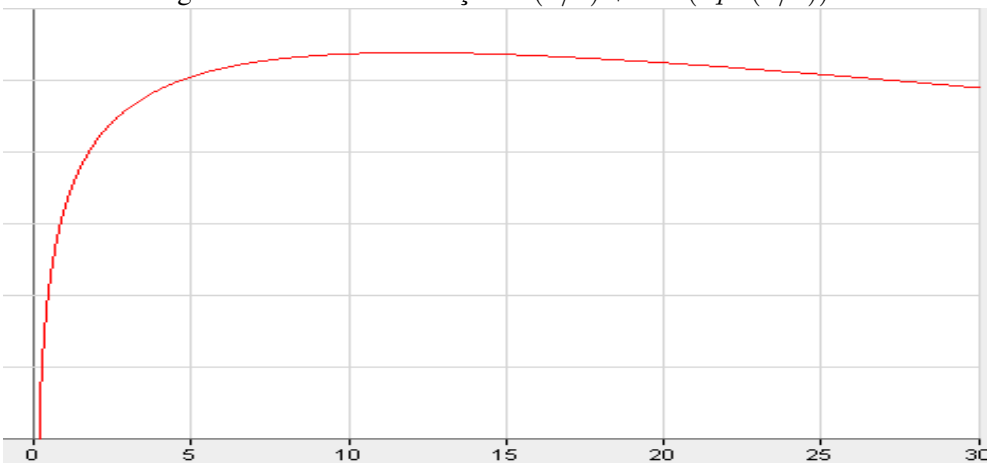
3.6.2 Integração do tempo de visualização

Para além da informação relativa às categorias e subcategorias, incluiu-se também a informação retirada a partir dos *Web Access Logs* relativa ao intervalo de tempo que um utilizador demorou na visualização de um item. Segundo [7], há uma relação entre o intervalo de tempo que um utilizador demora na visualização de determinado conteúdo e o interesse ou preferência por esse conteúdo. Neste artigo são utilizados dados resultantes de votação explícita para obtenção dos itens mais interessantes para o utilizador e posteriormente utilizam-se dados obtidos de forma implícita e o intervalo de tempo para obter os itens mais interessantes e constata-se que as previsões obtidas tendo em conta dados implícitos e o tempo gasto no conteúdo são quase tão precisas como utilizando dados obtidos de forma explícita. Perante isto verifica-se que a utilização do tempo gasto na visualização de determinado conteúdo poderá constituir uma informação de grande relevância.

Podemos obter o intervalo de tempo de visualização de determinada página referente a um item através da subtracção entre a referência temporal de visualização das distintas páginas pertencentes a determinada sessão (hora de acesso a uma página visitada depois do item menos hora de acesso à página do item). Para utilizar o intervalo de tempo no processo de previsão, para além de se utilizar directamente o valor da subtracção, também se considerou a existência de uma função que permita modelar a sua utilidade, possibilitando assim obter um valor tendo em conta o que o intervalo de tempo de visualização traduz em termos de preferência dos utilizadores, poucos segundos não traduzem preferência, minutos já traduzem maior preferência, mas muitos minutos traduzem alguma pausa na visualização da página e não propriamente preferência. Posto isto, foi considerado que até aos 15 segundos a utilidade é 0, isto porque o utilizador ou está em transição entre páginas ou verificou que não lhe interessa e mudou de página. A partir dos 15 segundos a função deverá ter um crescimento acentuado. Depois de alguns minutos a função deverá ter um comportamento decrescente. Para modelar o comportamento pretendido foram utilizadas duas funções $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$ e $ln(x/6) + 4 - (sqrt(x/3))$. Estas funções diferenciam-se através da rapidez com que atingem o valor máximo, uma demora sensivelmente 2 minutos ($3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$) e a outra 10 minutos ($ln(x/6) + 4 - (sqrt(x/3))$). Com esta diferença procura-se verificar qual da funções consegue modelar de forma mais aproximada as preferências dos utilizadores. Na primeira função 2 minutos de visualização corresponde ao máximo da preferência e na segunda aos 10 minutos. A função que melhores resultados obtiver será a que melhor consegue modelar as preferências do utilizador. Os gráficos das duas funções são apresentados nas figuras 3.6 e 3.7 respectivamente.

Figura 3.6: Gráfico da função $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$.

Figura 3.7: Gráfico da função $ln(x/6) + 4 - (sqrt(x/3))$.



Para a integração da informação relativa ao intervalo de tempo de visualização foi implementada a metodologia caracterizada pela substituição dos votos pelo valor/utilidade do tempo para o cálculo dos melhores vizinhos. No caso do utilizador activo o voto no item j ($v_{a,j}$) foi substituído pelo valor/utilidade do tempo despendido pelo utilizador activo na visualização do item j ($ut_{a,j}$). O mesmo foi aplicado ao nível dos outros utilizadores, em que $v_{i,j}$ é substituído por $ut_{i,j}$. A equação 3.1 apresenta a correlação após substituição dos votos pelo valor/utilidade do tempo. Esta operação também se aplica ao nível das variantes *Default Vote* e *Inverse User Frequency*. Na equação 2.1 procedeu-se à substituição do w pelo wut.

$$wut(a,i) = \frac{\sum_j (ut_{a,j} - \overline{ut}_a)(ut_{i,j} - \overline{ut}_i)}{\sqrt{\sum_j (ut_{a,j} - \overline{ut}_a)^2 \sum_j (ut_{i,j} - \overline{ut}_i)^2}} \tag{3.1}$$

Também se implementou uma metodologia que utiliza a soma dos pesos obtidos através dos votos e do tempo (wut + w) para calcular os melhores vizinhos. Realiza-se a soma de forma a impedir que quando um dos pesos for 0 o peso total seja 0.

A utilidade do tempo também foi aplicada aquando da determinação dos melhores itens a recomendar de três formas distintas:

- Quando $v_{i,j}$ é igual $\overline{v}_i$, verifica-se que $v_{i,j} - \overline{v}_i$ é 0 e no limite se todos os melhores vizinhos forem caracterizados por esta situação então consequentemente $p_{a,j} = \overline{v}_a$ (ver equação 2.1). Constata-se assim perda de informação como tal nestes casos substitui-se a informação relativa aos votos pelo valor/utilidade do tempo originando a seguinte equação:

$$p_{a,j} = \overline{v}_a + k \sum_{i=1}^{n} w(a,i) ut_{i,j} \tag{3.2}$$

- Utilizar o valor/utilidade do tempo em todos os casos ignorando a informação relativa aos votos (é apenas utilizada a equação 3.2). Este método visa obter o impacto que a informação do tempo tem sobre os resultados.

- Utilizar a soma do valor/utilidade do tempo e a informação dos votos ($v_{i,j} - \overline{v}_i$) com o intuito de determinar se constitui uma mais-valia em termos de resultados. Mais uma vez é utilizada a soma de forma preservar alguma informação quando uma das informações é nula.

3.6.3 Algoritmos a implementar

Nesta secção pretendem-se descrever os passos necessários para implementar os algoritmos e integração da informação apresentados na secção anterior.

O algoritmo que permite gerar recomendações para o utilizador activo, tem implícita a utilizando da correlação, *Default Vote* e *Inverse User Frequency*, com ou sem tempo. Em seguida são apresentados os passos considerados:

1. Obtém sessões do conjunto de teste que obedecem ao critério do "número de itens conhecidos";

   Para cada sessão de teste:

   (a) Obter sessões do conjunto de treino que visualizaram os itens conhecidos da sessão de teste;

   (b) Obter frequência dos itens conhecidos;

   (c) Chamar algoritmo que calcula melhores vizinhos (com maior peso) do utilizador activo.

   (d) Obter itens associados às sessões dos melhores vizinhos, excluindo os itens conhecidos.

   (e) Calcular $p_{a,j}$

   No caso do tempo, foram implementadas 3 abordagens distintas, nunca aplicadas em conjunto:

   i. Se $v_{i,j}$ igual a v_i, então o termo $(v_{i,j} - v_i)$ é substituído pelo valor/utilidade do tempo;

   ii. Utilizar apenas o valor/utilidade do tempo;

   iii. Utilizar a soma do valor/utilidade do tempo com a informação dos votos.

O algoritmo que possibilita o calculo dos melhores vizinhos (com maior peso) do utilizador activo, tem como objectivo a obtenção dos utilizadores mais semelhantes ao utilizador activo. Este algoritmo encontra-se organizado da seguinte forma:

1. Para cada sessão com itens iguais aos itens conhecidos do utilizador activo foram implementados 3 métodos alternativos:

   (a) Calcula o peso utilizando a correlação com ou sem *Inverse User Frequency*.

   (b) Calcula o peso utilizando a correlação com ou sem *Inverse User Frequency*, caso se verifique $v_{a,j} \neq \overline{v_{a,j}}$ e para os restantes casos utilizar o *Default Vote*.

   (c) Calcula o peso utilizando a correlação com ou sem *Inverse User Frequency*, caso se verifique $v_{a,j} \neq \overline{v_{a,j}}$ e/ou $v_{i,j} \neq \overline{v_{i,j}}$ e para os restantes casos utilizar o *Default Vote*.

2. Se o número de vizinhos a considerar já atingiu o limite máximo (só se pretendem obter os 10 melhores vizinhos), então proceder há remoção do vizinho com peso inferior ao peso do vizinho encontrado nesta iteração.

A escolha do método utilizado é feita explicitamente através de um parâmetro definido antes do cálculo dos melhores vizinhos.

O algoritmo que permite calcular as sugestões tendo em conta Subcategorias ou Categorias, visa calcular as recomendações utilizando a correlação, *Default Vote* e *Inverse User Frequency*, informação relativa aos itens, Categorias ou Subcategorias, com ou sem tempo. Os passos inerentes ao algoritmo são apresentados em seguida:

1. Obtém sessões do conjunto de teste que obedecem ao critério do "número de itens conhecidos".

   Para cada sessão de teste:

   (a) Obter as sessões do conjunto de treino que visualizaram os itens conhecidos da sessão de teste;

   (b) Obter frequência dos itens conhecidos;

   (c) Chamar algoritmo que calcula os melhores vizinhos do utilizador activo;

   (d) Obter sessão do conjunto de dados das Categorias ou Subcategorias com mesmo identificador da sessão de teste;

   Se existir sessão com o mesmo identificador:

   i. Obter Categorias ou Subcategorias associadas aos itens conhecidos.

   ii. Obter sessões do conjunto de dados das Categorias e Subcategorias, tendo em conta as sessões comuns ao conjunto de treino dos itens, de forma a excluir as sessões que não têm acesso a itens.

   iii. Obter frequência das Categorias ou Subcategorias identificadas no ponto (i). São consideradas apenas as sessões comuns ao conjunto de treino dos itens.

   iv. Calcular melhores vizinhos para o utilizador activo com os dados obtidos no ponto (ii) e (iii).

   v. Comparar vizinhos obtidos através dos itens com os vizinhos obtidos através das Categorias ou Subcategorias e escolher os melhores. Ou escolher os melhores 5 vizinhos de cada abordagem.

   (e) Obter itens associados às sessões dos melhores vizinhos;

   (f) Calcular $p_{a,j}$

   No caso do tempo, foram implementadas 3 abordagens distintas, nunca aplicadas em conjunto:

   i. Se $v_{i,j}$ igual a v_i, então o termo $(v_{i,j} - v_i)$ é substituído pelo valor/utilidade do tempo;

   ii. Utilizar apenas o valor/utilidade do tempo;

   iii. Utilizar a soma do valor/utilidade do tempo com a informação dos votos.

3.7 Análise exploratória

Nesta secção pretendem-se apresentar estatísticas relativas às variáveis consideradas nos diversos conjuntos de dados, com o propósito de caracterizar as variáveis, apresentando a frequência dos seus valores, possibilitando assim um melhor conhecimento e entendimento dos dados.

3.7.1 Análise dos dados relativos aos Itens

A análise apresentada em seguida é relativa ao conjunto de dados dos itens. O gráfico 3.8 apresenta os 12 itens mais visitados, de um total de 2570 itens visitados, o que corresponde a 23,3% dos itens existentes na base de dados (de notar que podem existir discrepâncias entre os itens na base de dados e os disponíveis no sítio).

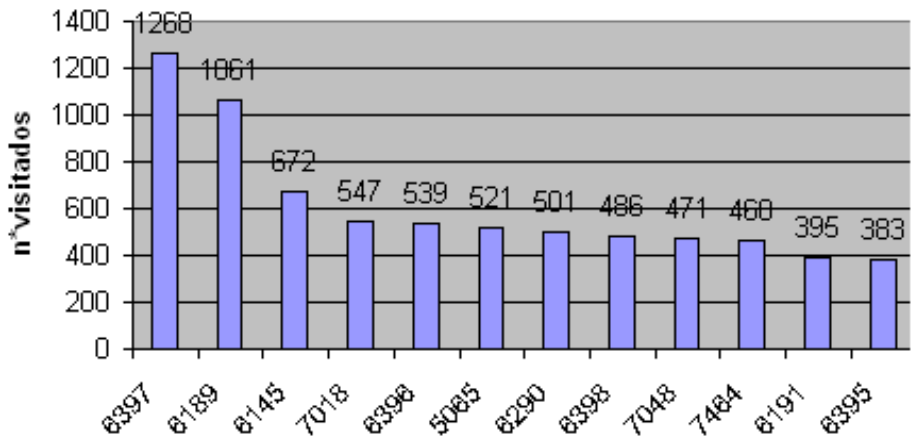


Figura 3.8: Número de acessos aos 12 itens mais visitados.

As descrições dos itens mais acedidos são:

- Portátil INMOVE WFL91 T7100;
- Computador SUPRIDES LOW-COST INTEL M71;
- Portátil INMOVE WH30 T5600;
- Portátil INMOVE WFL90-1 T7500;
- Portátil INMOVE WS96 T7500;
- Caixa Externa 2,5 USB2.0 P/DISCO IDE C/LEITOR Cartões =DIGIMATE;
- Portátil HP PAVILION MEDIA CENTER DV6559EA - CORE2DUO T7300;
- Portátil INMOVE WFL91 T7500;
- Portátil INMOVE WFT00 T7100;
- Portátil INMOVE WFL90-2 T7500 HDD160;
- Computador SUPRIDES M73;
- Portátil INMOVE WS96 T7100.

Ao analisar os itens mais visitados verifica-se que a maioria dos itens apresentados corresponde a portáteis e que os dois itens mais visitados apresentam uma grande diferença relativamente ao número de acessos, quando comparados com os restantes. O primeiro item tem sensivelmente o dobro do terceiro e o segundo aproximadamente o dobro do quarto.

O gráfico apresentado na figura 3.9, tem como intuito identificar qual a percentagem do número de visitas aos itens em sessões distintas (não são consideradas as visitas repetidas na mesma sessão). Isto é, 57% dos itens apresentam entre uma e quatro visitas em sessões distintas e assim sucessivamente.

O facto de a maioria dos itens ser visitado apenas entre 1 e 4 sessões distintas, pode afectar a qualidade das recomendações, quando a sessão do utilizador activo contemplar um destes itens, uma vez que ao existirem poucas sessões, o cálculo dos melhores vizinhos pode ser influênciado pela menor variabilidade, abrangendo poucos cenários de interesse dos utilizadores.

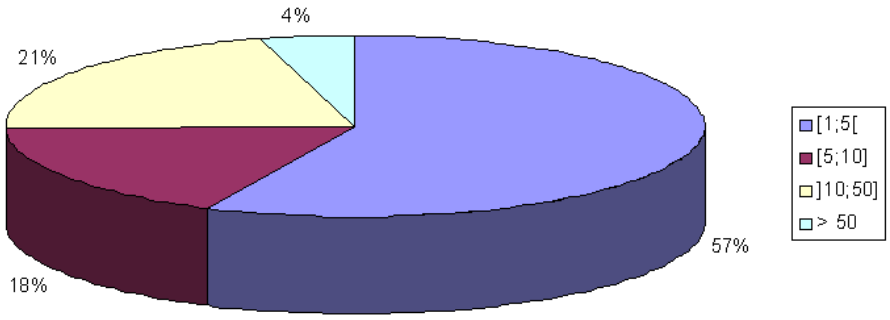


Figura 3.9: Histograma da frequência de visitas aos itens em intervalo de valores.

O gráfico apresentado na figura 3.10, tem como propósito identificar qual a percentagem do número visitas a itens, por sessão, em intervalo de valores. Ou seja, 72% das sessões apresentam 1 visita a itens e assim sucessivamente. O número médio de itens visualizados por sessão é de 1,7. O facto de a maioria das sessões apenas apresentar uma visita a um item é relevante, uma vez que pode ser insuficiente para realizar uma recomendação de qualidade para estas sessões, isto porque a escassa informação conhecida acerca do utilizador dificulta a determinação dos seus interesses.

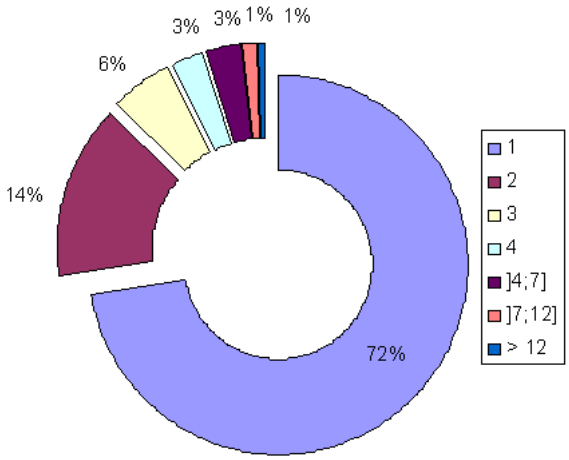


Figura 3.10: Histograma da frequência de visitas a itens, por sessão.

3.7.2 Análise dos dados relativos às Categorias

A análise apresentada em seguida é relativa ao conjunto de dados das categorias. A figura 3.11 apresenta o gráfico relativo à percentagem do número de visitas às categorias (62,5% das categorias apresentam visitas). Com este pretende-se mostrar qual a distribuição de visitas por categoria.

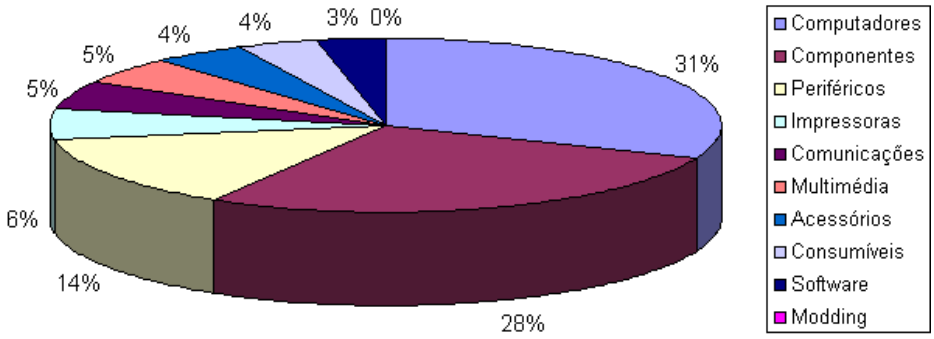


Figura 3.11: Percentagem do número de visitas a cada categorias.

O gráfico apresentado na figura 3.12, tem como propósito identificar qual a percentagem do número visitas a categorias, por sessão, em intervalo de valores. Através deste pode observar-se que 84% das sessões apresentam uma visita a categorias e assim sucessivamente. Neste caso constata-se que o número de visitas a categorias por sessão é bastante diminuto. O número médio de categorias visualizadas por sessão é de 1,2. O facto de a maioria das sessões apenas apresentar uma visita a uma categoria é relevante, uma vez que pode ser insuficiente para realizar uma recomendação de qualidade para estas sessões, isto porque a escassa informação conhecida acerca do utilizador dificulta a determinação dos seus interesses.

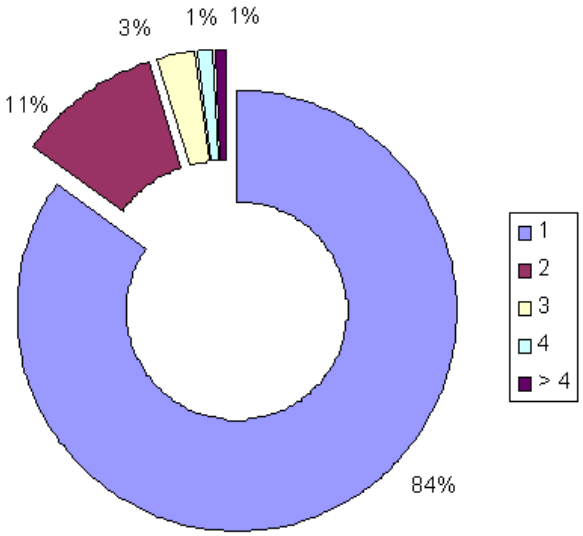


Figura 3.12: Histograma da frequência de visitas a categorias, por sessão.

3.7.3 Análise dos dados relativos às Subcategorias

A análise apresentada em seguida é relativa ao conjunto de dados das subcategorias. A figura 3.13 apresenta o gráfico relativo às 20 subcategorias mais visitadas, de um total de 187 subcategorias visitadas, o que representa 82,4% das subcategorias existentes. Através do gráfico constata-se que a subcategoria mais visitada apresenta uma quantidade de acessos bastante superior às outras subcategorias.

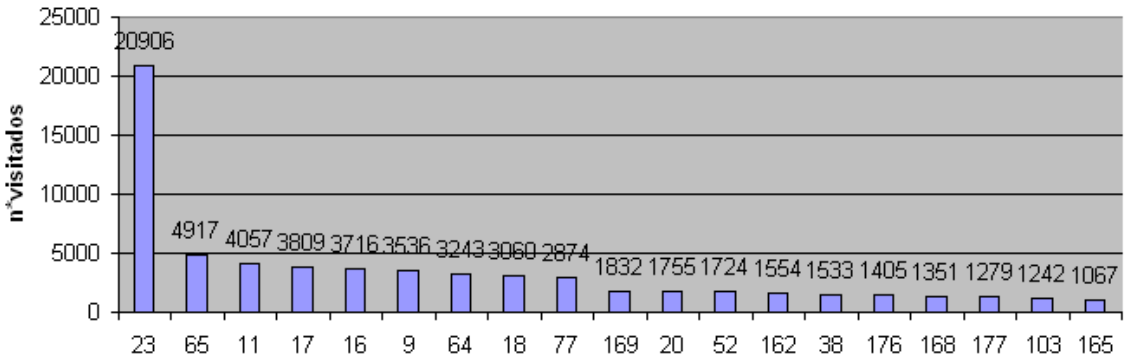


Figura 3.13: Número de acessos às 20 subcategorias mais visitadas.

As descrições das subcategorias apresentadas no gráfico anterior são:

- Portáteis;
- Monitores;
- Discos;
- Placas Gráficas;
- *Motherboards*;
- Caixas;
- Armazenamento;
- Processadores;
- Computador Pessoal (PC);
- Câmaras Fotográficas;
- *Barebones* (Mini PC);
- Multifunções;
- Memórias (PC);
- Tinteiros;
- PDA/SMARTPHONE;

- Redes S/ Fios;
- Plotters;
- *Upgrade*;
- Cartões de Memória;

O gráfico apresentado na figura 3.14, tem como intuito identificar a percentagem de visitas às subcategorias, para cada intervalo de valores, em sessões distintas (não são consideradas as visitas repetidas na mesma sessão). Pode então observar-se que 34% das subcategorias têm entre 10 e 50 visitas em sessões distintas. É de salientar que, ao contrário do observado para os itens, a maior percentagem de subcategorias encontram-se nos intervalos mais elevados, ou seja, têm mais de 10 visitas. Este facto permite melhorar a qualidade das recomendações, quando a sessão do utilizador activo contemplar uma destas categorias, uma vez que ao existirem muitas sessões, o cálculo dos melhores vizinhos pode ser afectado pela maior variabilidade, abrangendo muitos cenários de interesse dos utilizadores.

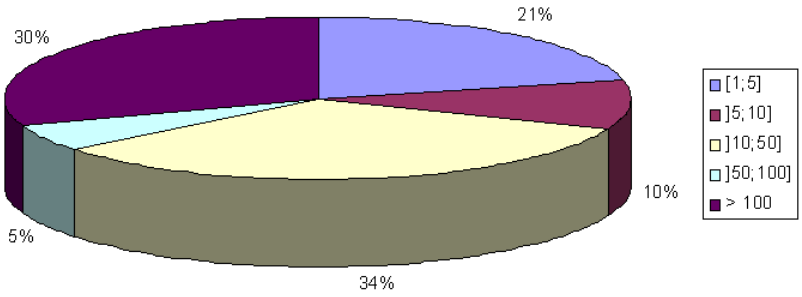


Figura 3.14: Histograma da frequência de visitas às subcategorias por intervalos de valores.

Na figura 3.15 é apresentado o gráfico relativo à percentagem do número visitas a subcategorias, por sessão. Isto é, 74% das sessões apresentam uma visita. A média de visitas a subcategorias por sessão é de 1,7. O facto de maioria das sessões apenas apresentar uma visita a uma subcategoria é relevante, uma vez que pode ser insuficiente para realizar uma recomendação de qualidade para estas sessões, isto porque a escassa informação conhecida acerca do utilizador dificulta a determinação dos seus interesses.

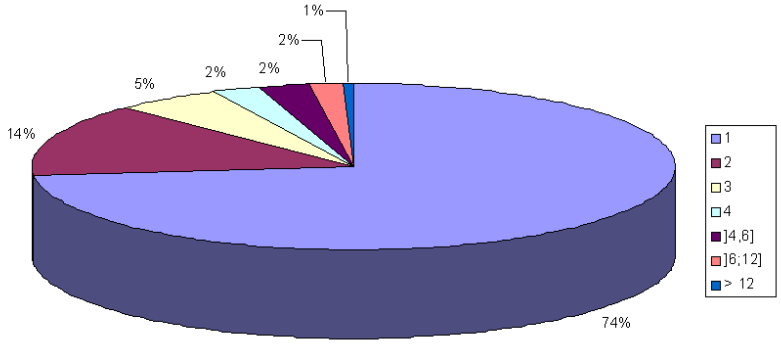


Figura 3.15: Histograma da frequência de visitas a subcategorias, por sessão.

3.7.4 Análise dos dados relativos ao Tempo

A análise apresentada em seguida é relativa ao conjunto de dados do tempo. A figura 3.16 apresenta um gráfico com as percentagens de visitas a itens tendo em conta o intervalo de tempo de visualização. Ou seja, em 48% das visitas a itens, os utilizadores demoraram entre 0 e 15 segundos na sua visualização, depois temos o intervalo entre 15 segundos e 1 minuto e posteriormente são utilizados intervalos de 2 minutos.

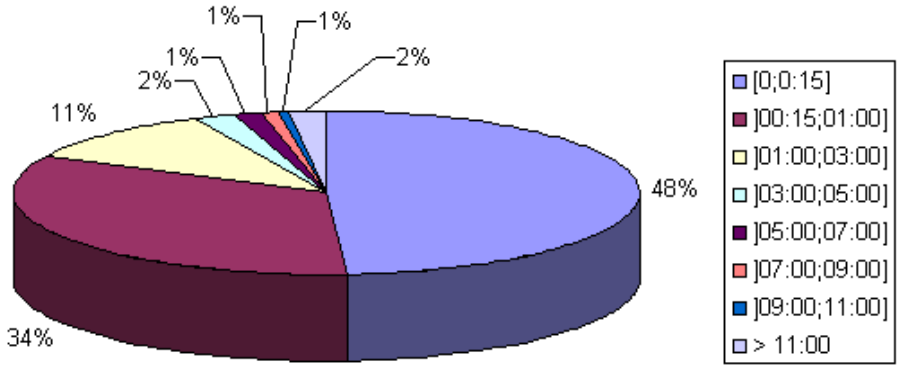


Figura 3.16: Histograma da frequência de visitas a itens por intervalo de tempo de visualização.

O gráfico apresentado na figura 3.17 visa identificar a percentagem de sessões por intervalo de tempo de visualização. Através deste pode constatar-se que 84% das sessões apresentam visitas a itens por um periodo de tempo conpreendido entre 0 e 2 minutos e assim sucessivamente. A média do tempo de visualização por sessão é de 1 minuto e 40 segundos.

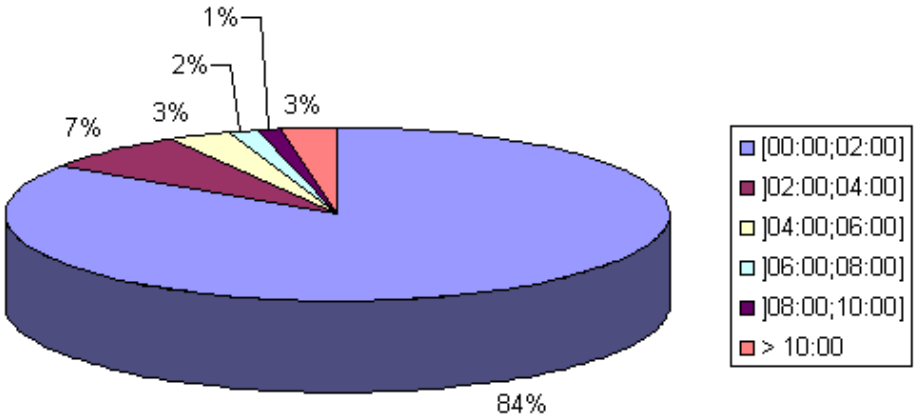


Figura 3.17: Histograma da frequência de sessões por intervalo de tempo de visualização.

3.7.5 Sumário da análise dos dados

Perante o que foi dito nas secções anteriores pode constatar-se que para os itens, subcategorias e categorias a maioria das sessões apresenta apenas 1 acesso, o que pode comprometer a qualidade das recomendação para estas sessões, porque a escassa informação conhecida acerca do utilizador dificulta a determinação dos seus interesses.

Verifica-se ainda que os itens na sua maioria são visualizados entre uma e quatro sessões, o que pode afectar a qualidade das recomendações, quando a sessão do utilizador activo contemplar um destes itens, uma vez que ao existirem poucas sessões, o cálculo dos melhores vizinhos pode ser afectado pela menor variabilidade, abrangendo poucos cenários de interesse dos utilizadores. Ao contrário para as subcategoria e categorias na sua maioria são visualizados mais de 10 vezes e mais de 91 vezes respectivamente, podendo assim possibilitar recomendações de qualidade, uma vez que ao existirem muitas sessões, o cálculo dos melhores vizinhos pode ser afectado pela maior variabilidade, abrangendo muitos cenários de interesse dos utilizadores. Estas diferenças estão relacionadas com a diversidade de itens (2570), subcategoria (187) e categorias (10) visitadas, menor diversidade permite maior número de acessos. Esta última constatação possibilita apoiar o interesse que pode advir da integração no *CF* da informação relativa às categorias e subcategorias.

Capítulo 4

Experiências e Resultados

4.1 Metodologia experimental

Com o intuito de levar a cabo a obtenção de resultados inerentes à aplicação dos algoritmos e abordagens anteriormente referidas, procedeu-se à utilização da seguinte metodologia: 80% dos dados como treino e os restantes 20% como conjunto de teste. A tabela recuperada da secção 2.4.1, permite exemplificar gráficamente a separação dos dados.

		j_1	..	j_l	..	j_{l+2}	j_k
Conj. treino	i_1						
	i_2	$v_{2,1}$					
	:						
Conj. teste	i_a						
	i_q						
	i_m						

Tabela 4.1: Esquema da organização conceptual do conjunto de treino e teste.

A metodologia de avaliação utilizada foi apresentada no secção 2.4.1 e pressupõe a divisão em duas partes das sessões de teste, sendo uma das partes constituída por itens conhecidos (de j_1 até j_l), a outra parte é composta pelos restantes itens, que são utilizados para realizar a avaliação da previsão (itens escondidos). Estes itens em termos de avaliação são considerados como relevantes (ver 2.4). Para a execução das experiências foram utilizados 4 parâmetros experimentais distintos caracterizados pelo facto de j_l ser igual 2, 5 e 10 itens e o parâmetro "All but 1" que corresponde à última linha apresentada em 4.1.

Em cada parâmetro experimental, os 20% do conjunto de teste representam um número diferente de sessões utilizadas para previsão de itens, pois o número de sessões que preenche os requisitos é distinto. Para o caso do conjunto de dados referente aos itens, quando dados como conhecidos 2 itens, o número de sessões passiveis de previsão é de 253; dados 5 itens temos 66; dados 10 itens temos 24; e utilizando o parâmetro "All but 1" temos 1017. É de salientar que nos três primeiros parâmetros é estabelecido como requisito para além do número de itens conhecidos, a exitência de pelo menos 2 itens escondidos, de forma a garantir que há pelo menos 2 itens considerados relevantes aquando da avaliação. Por exemplo quando conhecidos dois itens as sessões escolhidas têm no mínimo 4 itens.

Depois de gerados os itens previstos para cada sessão de teste, tendo por base as sessões do conjunto de treino, são calculadas para cada sessão de teste as medidas de avaliação *Precision*, *Recall* e *F1*,

descritas na secção 2.4. Posteriormente, é calculada a média dos resultados das medidas de avaliação de todas as sessões de teste de modo a obter um resultado único, com a finalidade de facilitar a análise.

Os resultados obtidos são apresentados numa tabela com a seguinte organização: na primeira coluna são identificados os parâmetros utilizados; na segunda, terceira e quarta coluna a média da *Precision*, *Recall* e *F1* respectivamente. A última coluna destina-se a apresentar o número de sessões para as quais foram geradas previsões. Podem surgir casos em que o `Sistema de Recomendação` não gera previsões, porque devido há insuficiência de dados no conjunto de treino não são encontrados utilizadores com os mesmos itens visitados que o utilizador activo.

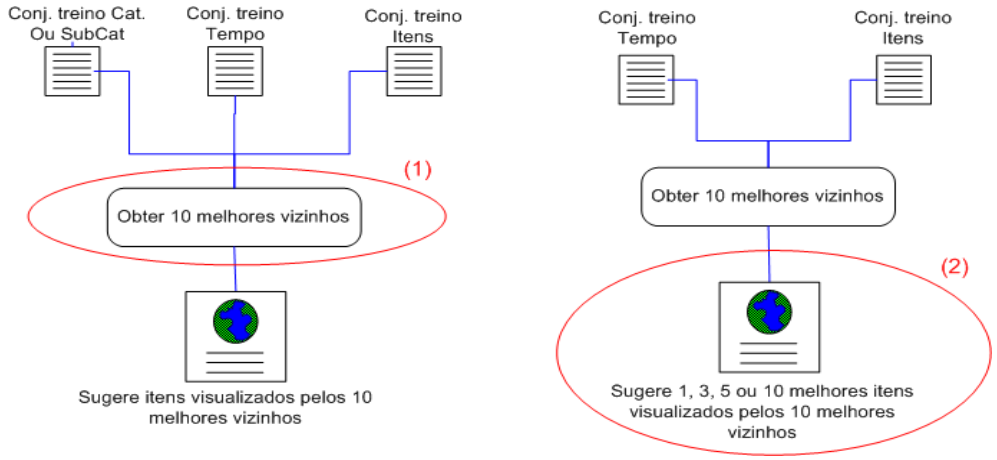
As secções 4.2, 4.3 e 4.4 vão incidir sobre a fase de escolha dos melhores vizinhos, que corresponde à área (1) identificada no diagrama 4.1. Para esta fase propõem-se as seguintes hipóteses:

- Comparar os resultados do algoritmo de correlação isoladamente e em conjunto com as suas variantes *Inverse User Frequency* e *Default Vote* com o intuito de verificar qual deles consegue alcançar melhores resultados;
- Comparar a utilização do conjunto de dados das subcategorias ou categorias isoladamente e em conjunto com a informação relativa aos itens com o propósito de verificar qual deles consegue alcançar melhores resultados;
- Comparar a utilização do conjunto de dados dos itens e a informação do intervalo de tempo de visualização sem utilidade associada, com a utilização das duas funções distintas de utilidade, com o objectivo de verificar qual deles consegue alcançar melhores resultados.

Relativamente à secção 4.6 tem como intuito avaliar as metodologias aplicadas ao nível da escolha dos itens com maior valor previsto (área (2)), uma vez que o número de itens está limitado ao contrário do que aconteceu nas hipóteses apresentadas anteriormente. Para esta fase propõem-se a seguinte hipótese:

- Comparar os resultados obtidos utilizando apenas informação relativa aos itens (votos) e os resultados obtidos utilizando a informação dos itens em conjunto com a informação relativa ao intervalo de tempo de visualização ou utilizando apenas informação relativa ao tempo.

Figura 4.1: Diagrama representativo das fases de *Collaborative Filtering*.



4.2 Comparação dos algoritmos

Nesta secção pretende-se aplicar o algoritmo de correlação isoladamente e em conjunto com as suas variantes *Inverse User Frequency* e *Default Vote* com o intuito de comparar os resultados e verificar qual deles consegue alcançar melhores resultados, utilizando para tal a metodologia apresentada na secção anterior.

Como primeira abordagem para o problema seguiu-se a implementação descrita no capítulo 2, onde se propõe a utilização de algoritmos *Memory-Based*, que são caracterizados pela utilização de uma base de dados com o histórico de votos em itens, a partir da qual se procuram gerar recomendações para os utilizadores do sítio.

Os resultados apresentados na tabela 4.2 foram obtidos a partir da implementação do algoritmo de *Correlação* (abreviatura *Corr*), descrito em detalhe na secção 2.2.2. Nesta tabela também são apresentados os resultados obtidos a partir da implementação de uma das variantes designada por *Inverse User Frequency* (abreviatura *CIF*), que é oriunda da área de "Information retrieval" e pode ser transposta para o *Collaborative Filtering*, considerando que os itens visitados com maior frequência têm menos peso no calculo da semelhança entre utilizadores que os itens menos visitados. Para uma descrição mais detalhada, ver secção 2.2.2.

Depois de uma análise mais cuidada da forma de obtenção dos resultados verificou-se que, em muitas sessões a média dos votos ($\overline{v}_j$) é igual ao voto (v_j) o que leva a que um dos factores da formula 2.2 seja 0 e consequentemente o peso (w) também seja. Se esta situação se verificar para todos os utilizadores que visitaram os mesmos itens que o utilizador activo, temos $p_{a,j} = \overline{v}_a$, ocorrendo assim perda de informação.

Para tentar evitar esta situação foi implementada a abordagem *Default Vote*, apresentada na secção 2.2.2.

O *Default Vote* é uma variante do algoritmo de correlação e é especialmente indicado para casos em que existem poucos itens visitados quer ao nível do utilizador activo, quer ao nível dos utilizadores do histórico.

Para aplicação do método referido anteriormente foram considerados dois cenários:

- Utilizar quando a sessão do utilizador activo apresenta $v_{a,j}$ igual a $\overline{v}_{a,j}$ (abreviatura *Cact*);
- Utilizar quando a sessão do utilizador activo apresenta $v_{a,j}$ igual a $\overline{v}_{a,j}$, e/ou quando o outro utilizador apresenta $v_{i,j}$ igual a $\overline{v}_{i,j}$ (abreviatura *Cout*).

Os resultados inerentes a estes dois métodos podem-se também visualizar na tabela 4.2.

Parâmetros	AVG(precision)(%)				AVG(recall)(%)				AVG(f1)(%)				Sessões com Prev. (%)			
	Cor	*CIF*	*Cact*	*Cout*	*Cor*	*CIF*	*Cact*	*Cout*	*Cor*	*CIF*	*Cact*	*Cout*	*Cor*	*CIF*	*Cact*	*Cout*
2 itens	5,3	6,4	6,4	9,7	39,1	30,7	30,2	24,9	9,3	10,6	10,6	14,0	98,4	97,6	97,2	89,7
5 itens	4,5	5,1	5,1	7,8	30,7	20,6	20,6	11,4	7,8	8,2	8,2	9,3	97,0	97,0	97,0	90,9
10 itens	3,9	4,0	3,7	5,3	25,7	11,0	8,9	6,6	6,8	5,9	5,2	5,9	100,0	100,0	100,0	95,8
"All but 1"	3,6	4,6	4,6	5,0	49,2	45,5	45,5	39,0	6,7	8,4	8,4	8,9	93,8	92,9	92,8	86,9

Tabela 4.2: Resultados obtidos a partir da implementação do algoritmo de *Correlação* e suas variantes.

Em seguida procede-se à análise dos resultados com intuito de verificar se é valida a hipótese formulada no inicio desta secção: aplicar o algoritmo de correlação isoladamente e em conjunto com as suas variantes *Inverse User Frequency* e *Default Vote* com o intuito de verificar qual deles consegue alcançar melhores resultados. Esta hipótese foi subdividida em quatro observações principais mais especificas com o intuito de facilitar a análise.

> *Observação 1*: A implementação da *Correlação* com a variante *Inverse User Frequency* (*CIF*) apresenta melhores resultados relativamente à *Correlação* (*Corr*).

Após comparação dos resultados verifica-se uma melhoria dos resultados da *CIF* relativamente à *Corr*, caracterizada pela existência de valores mais elevados ao nível da métrica *Precision* e *F1*, exceptuando para esta última o parâmetro "10 itens", onde se constata que a diferença dos valores do *Recall* entre os dois métodos apresenta o seu maior valor e a diferença do valor da *Precision* o menor de todos os parâmetros aplicados. Para a métrica *Recall* a *Correlação* apresenta os resultados mais elevados. Estes resultados indiciam que com a *Correlação* ocorre a sugestão de maior número de itens prejudicando a *Precision*, mas aumentando a *Recall*.

No que diz respeito à percentagem de sessões com previsão constata-se uma ligeira diminuição quando utilizada a variante (*CIF*) e os parâmetros "Dados 2 itens" e "All but 1". Pode verificar-se que esta situação coincide com o facto de a diferença do valor da *Precision* entre os dois métodos ser maior que nos outros parâmetros, apontado para uma relação entre os factos.

Tal como esperado verifica-se que o valor de *F1* da *CIF* é maior em todos os parâmetros, com excepção do "10 itens" devido ao elevado valor da métrica *Recall*. Este parâmetro é caracterizado pelo reduzido número de sessões teste (24), pelo que com um conjunto de dados com mais sessões que encaixassem nos requisitos deste parâmetro talvez os resultados obtidos estivessem de acordo com os restantes parâmetros.

> *Observação 2*: A implementação da *Correlação* com as variantes *Inverse User Frequency* e *Default Vote* no segundo cenário (*Cout*) apresenta melhores resultados relativamente à *Correlação* com as variantes *Inverse User Frequency* e *Default Vote* no primeiro cenário (*Cact*).

Relativamente à utilização da variante *Default Vote* verificaram-se resultados mais elevados no segundo cenário (*Cout*) em todos os parâmetros, no que diz respeito às medidas *F1* e *Precision*, ocorrendo a situação inversa ao nível da medida *Recall*. Constata-se também uma situação semelhante à relatada na observação anterior, pois pode observar-se uma relação entre a diminuição do número de sessões com previsão e a *Precision* e *F1*. No primeiro cenário (*Cact*) a média de sessões com previsão (considerando todos os parâmetros) é de 97%, enquanto que no segundo cenário é de 91%.

Como esperado a utilização do segundo cenário (*Cout*) conduziu a melhores resultados, pois diminui a perda de informação gerada pelo facto de $v_{a,j}$ igual a $\overline{v}_{a,j}$ para o utilizador activo e/ou quando o outro utilizador apresenta $v_{i,j}$ igual a $\overline{v}_{i,j}$.

> *Observação 3*: A implementação da *Correlação* com as variantes *Inverse User Frequency* e *Default Vote* no primeiro cenário (*Cact*) apresenta melhores resultados relativamente à *Correlação* com a variante *Inverse User Frequency* (*CIF*).

Ao comparar *Cact* com os resultados da implementação da *Correlação* com a variante *Inverse User Frequency* verifica-se grande semelhança, pois apenas há diferenças ao nível do parâmetro "10 itens",

onde se constata uma redução dos valores obtidos para todas as métricas. Relativamente às sessões com previsão pode observar-se que o valor médio de *Cact* (96,8) é inferior ao valor de *CIF* (96,9), embora de forma pouco significativa.

Ao contrário do esperado os resultados são semelhantes, pois previa-se que o facto de *Cact* ter em conta que $v_{a,j}$ igual a $\overline{v}_{a,j}$ pudesse constituir ganho de informação que se traduzisse em melhores resultados. Também se verifica um impacto negativo sobre o parâmetro "10 itens".

> *Observação 4*: A implementação da *Correlação* com as variantes *Inverse User Frequency* e *Default Vote* no segundo cenário (*Cout*) apresenta melhores resultados relativamente à *Correlação* com a variante *Inverse User Frequency* (*CIF*).

Ao fazer a comparação dos resultados obtidos na implementação das abordagens referidas nesta observação, verifica-se que os resultados de *Cout* são mais elevados nas métricas *Precision* e *F1*. Verificam-se também maiores valores ao nível da métrica do *Recall* para *CIF*, o que é indicativo que realiza a previsão de um maior número de itens que o *Cout*. Relativamente ao número de sessões com previsão pode observar-se uma diminuição em todos os parâmetros ao nível de *Cout*. Tal como nas hipóteses anteriores constata-se uma relação entre a diminuição do número de sessões com previsão, *Precision* e *F1*.

Como esperado verificou-se uma melhoria dos resultados, confirmando que a estratégia seguida em *Cout* conduziu a ganho de informação.

A tabela 4.3 visa mostrar os resultados mais elevados obtidos nesta secção para cada métrica e parâmetro, identificando também qual a abordagem e percentagem de sessões com previsão inerente. Idêntica análise se pode fazer através dos gráficos apresentados em 4.2 e 4.3. A abordagem que mais contribui para esta tabela é a Correlação, *Inverse User Frequency* e *Default Vote* para o utilizador activo e utilizador a comparar. A utilização da Correlação isoladamente tem um bom desempenho na métrica *Recall*, assim como na métrica *F1* e parâmetro "Dados 10 itens". Este resultado é bastante surpreendente dados os restantes resultados ao nível desta métrica, no entanto é de salientar o elevado valor do *Recall* para esta abordagem e parâmetro que acaba por contribuir de forma decisiva para o valor de *F1*. Talvez o facto do reduzido número de sessões de teste (24) contribua para este resultado inesperado.

Tendo em conta a hipótese colocada no início desta secção e toda a análise apresentada, constata-se que tal como esperado as abordagens propostas que utilizam a correlação em conjunto com as variantes *Default Vote* e *Inverse User Frequency*, com especial evidência da *Cout*, conseguiram globalmente obter melhores resultados com a excepção da métrica *Recall* onde *Corr* apresenta melhores resultados em todos os parâmetros.

	Dados 2 itens	**Dados 5 itens**	**Dados 10 itens**	**'All but 1'**
AVG(precision)(%)	9,7	7,8	5,3	5,0
Sessões com Prev. (%)	89,7	90,9	95,8	86,9
Abordagem	*Cout*	*Cout*	*Cout*	*Cout*
AVG(recall)(%)	39,1	30,7	25,7	49,2
Sessões com Prev. (%)	98,4	97,0	100	93,8
Abordagem	*Corr*	*Corr*	*Corr*	*Corr*
AVG(f1)(%)	11,2	8,0	6,3	7,8
Sessões com Prev. (%)	89,7	90,9	100	86,9
Abordagem	*Cout*	*Cout*	*Corr*	*Cout*

Tabela 4.3: Melhores resultados obtidos na secção 4.2 e respectiva abordagem onde foi apresentado o resultado.

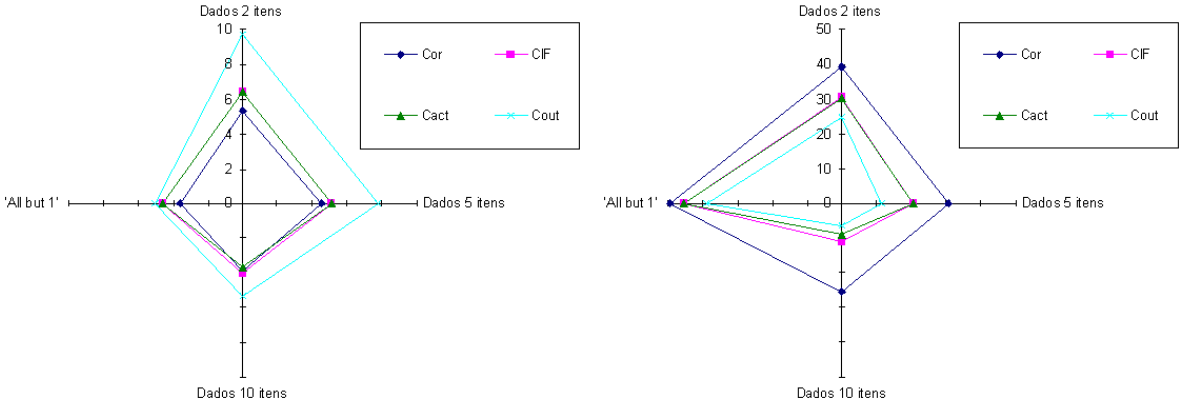


Figura 4.2: Comparação gráfica das várias abordagens e parâmetros para as medidas de avaliação *Precision* (gráfico à esquerda); e *Recall*.

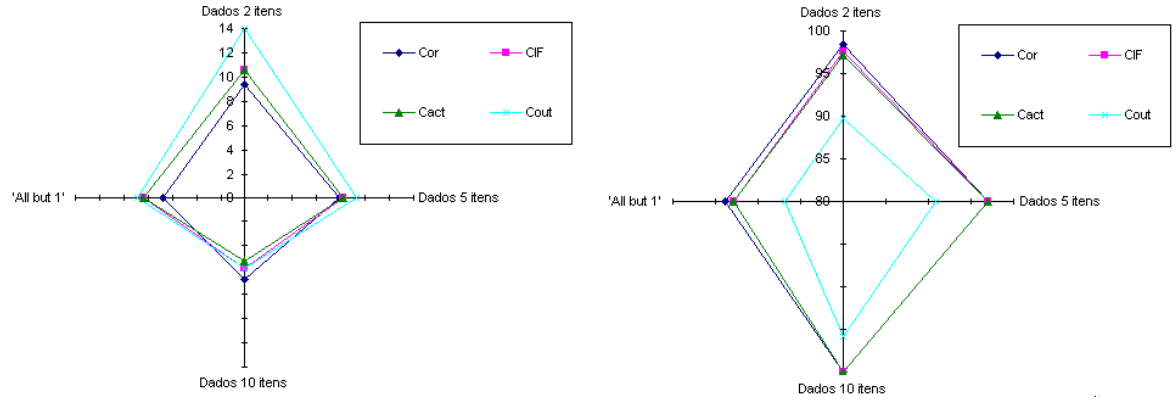


Figura 4.3: Comparação gráfica das várias abordagens e parâmetros para a medida de avaliação *F1* (gráfico à esquerda); e percentagem de sessões com previsão.

4.3 Integração de informação relativa às Categorias e Subcategorias

Esta secção tem como intuito aplicar o método apresentado em 3.6.1, o qual pretende integrar informação relativa às subcategorias e categorias, quer ao nível da relação existente com os itens, quer ao nível dos acessos registados nos *Web Access Logs*. Pretende-se assim utilizar o conjunto de dados das subcategorias ou categorias isoladamente e em conjunto com a informação relativa aos itens com o propósito de comparar os resultados e verificar qual deles consegue alcançar melhores resultados.

A utilização desta informação tem como base uma das abordagens implementadas na secção anterior, mais concretamente a última abordagem caracterizada pela Utilização da Correlação, *Inverse User Frequency* e *Default Vote* para o utilizador activo e utilizador a comparar (*Cout*).

Como primeira abordagem procedeu-se à utilização apenas da informação relativa às categorias ou subcategorias para obter previsões dos itens. Nesta situação são utilizadas as categorias ou subcategorias em que se enquadram os itens conhecidos, calculados os melhores vizinhos utilizando o conjunto de dados das categorias ou subcategorias, tendo em conta o facto de só serem utilizadas as sessões comuns com o conjunto de treino dos itens. Depois de determinados os melhores vizinhos obtêm-se os itens visitados por estes, com a condição de serem diferentes dos itens conhecidos e calculam-se os valores previstos. Os resultados obtidos referentes à utilização das subcategorias e categorias são apresentados nas colunas "*Sub*" e "*Cat*" das tabelas 4.4 e 4.5 respectivamente.

Para além da abordagem apresentada anteriormente, procedeu-se à utilização da abordagem descrita em 3.6.1, onde se propõe a conjugação do conjunto de dados relativo aos itens e o conjunto de dados das categorias ou subcategorias, para determinar os melhores vizinhos. Os resultados obtidos referentes à utilização das subcategorias e categoria são apresentados nas colunas "*Sit*" e "*Cit*" das tabelas 4.4 e 4.5 respectivamente.

Também se utilizou a abordagem apresentada em 3.6.1 caracterizada pela escolha dos 5 melhores vizinhos de cada conjunto de dados. Os resultados obtidos referentes à utilização das subcategorias e categoria são apresentados nas colunas "*Sit5*" e "*Cit5*" das tabelas 4.4 e 4.5 respectivamente.

Em seguida vai realizar-se a análise de resultados com intuito de verificar se é valida a hipótese formulada no inicio desta secção. Esta hipótese foi subdividida em quatro observações principais mais especificas com o intuito de facilitar a análise.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)			Sessões com Prev. (%)		
	Sub	*Sit*	*Sit5*	*Sub*	*Sit*	*Sit5*	*Sub*	*Sit*	*Sit5*	*Sub*	*Sit*	*Sit5*
2 itens	14,3	12,4	11,6	19,0	21,8	20,4	16,3	15,8	14,8	24,9	39,1	78,3
5 itens	8,3	5,4	7,2	9,4	7,7	9,5	8,8	6,3	8,2	34,8	43,9	75,8
10 itens	4,8	4,2	4,6	7,1	6,1	3,8	5,7	5,0	4,2	29,2	45,8	79,2
"All but 1"	3,5	4,8	5,7	21,3	35,9	32,2	6,0	8,5	9,7	21,2	41,9	70,6

Tabela 4.4: Resultados obtidos a partir da integração de informação relativa às Subcategorias.

> *Observação 1*: A implementação da abordagem em que se utilizam o conjunto de dados dos itens e das subcategorias (*Sit*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas o conjunto de treino das subcategorias (*Sub*).

Ao comparar estas duas abordagens constata-se que *Sub* apresenta melhores resultados em todos os parâmetros e métricas excepto ao nível do parâmetro "2 itens" para a métrica *Recall* e do parâmetro "All but 1", onde se verifica um cenário inverso tal que *Sit* é melhor em todas as métricas. É de

salientar o facto de que a média de sessões com previsão para *Sub* ser de 27,5% o que constitui um valor substancialmente inferior relativamente a *Sit* que apresenta um valor de 42,7%.

Os resultados obtidos não estão de acordo com o esperado pois esperava-se que a inclusão da informação relativa aos itens permitisse uma melhoria em todos os parâmetros e não só num como se verificou. No entanto a inclusão da informação relativa aos itens possibilitou sem dúvida o aumento do número de sessões com previsão. O resultado obtido em "All but 1" indicia que a abordagem pode ter melhores resultados quando existir um histórico de maior dimensão e maior informação disponível acerca dos itens visitados pelo utilizador.

> *Observação 2*: A implementação da abordagem em que se utilizam os 5 melhores vizinhos obtidos a partir do conjunto de dados dos itens e das subcategorias (*Sit5*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas o conjunto de treino das subcategorias (*Sub*).

Ao comparar os resultados obtidos nas duas abordagens referidas na observação verifica-se que *Sit5* apresenta melhores resultados para todas as métricas no parâmetro "All but 1" e também para a métrica *Recall* com excepção de "5 itens". Ao nível das sessões com previsão constata-se uma diferença significativa, pois a média de *Sub* fica pelos 27,5% enquanto *Sit5* fica pelos 76%.

Os resultados não estão de acordo com o esperado pois verifica-se que a inclusão da informação relativa aos itens não permitiu uma melhoria em todos os parâmetros. No entanto permitiu melhores resultados ao nível da métrica *Recall* e possibilitou sem dúvida o aumento do número de sessões com previsão. O resultado obtido em "All but 1" indicia que a abordagem *Sit5* pode ter melhores resultados quando existir um histórico de maior dimensão e mais informação disponível acerca dos itens visitados pelo utilizador.

Ao comparar os resultados obtidos em *Sit5* e *Sit* observa-se que *Sit5* apresenta melhores resultados na métrica *Precision* em "5 itens", "10 itens" e "All but 1". Na métrica *Recall* apenas em "5 itens" e na métrica *F1* em "5 itens" e "All but 1". Verifica-se assim que *Sit5* é melhor em todas as métricas no parâmetro 5 itens. Ao nível das sessões com previsão constata-se uma diferença significativa pois a média de *Sit* fica pelos 42,7% enquanto *Sit5* fica pelos 76%.

É de salientar que se verifica uma relação clara entre a utilização do conjunto de dados relativo aos itens e o aumento da percentagem das sessões com previsão.

Em baixo é apresentada a análise de resultados obtidos com a utilização do conjunto de dados das categorias.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)			Sessões com Prev. (%)		
	Cat	*Cit*	*Cit5*	*Cat*	*Cit*	*Cit5*	*Cat*	*Cit*	*Cit5*	*Cat*	*Cit*	*Cit5*
2 itens	7,5	8,3	11,4	18,4	22,9	19,2	10,7	12,2	14,3	12,3	24,1	74,7
5 itens	2,7	2,1	8,8	8,8	5,7	9,1	4,1	3,1	8,9	13,6	22,7	71,2
10 itens	3,3	3,7	4,2	5,0	5,2	1,3	4,0	4,3	2,0	20,8	33,3	70,8
"All but 1"	3,3	5,6	5,1	26,6	44,1	30,6	5,9	9,9	8,7	12,2	29,0	69,0

Tabela 4.5: Resultados obtidos a partir da integração de informação relativa às Categorias.

> *Observação 3*: A implementação da abordagem em que se utilizam o conjunto de dados dos itens e das categorias (*Cit*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas o conjunto de treino das categorias (*Cat*).

Comparando os resultados obtidos nas duas abordagens referidas na observação constata-se que *Cit* apresenta melhores resultados em todas as métricas e parâmetros com excepção do parâmetro "5 itens", onde se podem observar piores resultados em todas as métricas. Ao nível da percentagem de sessões com previsão verifica-se a existência de maior média em *Cit* com 27,3%, enquanto que *Cat* apresenta apenas uma média de 14,7%.

Como previsto *Cit* globalmente apresenta melhores resultados que *Cat* permitindo constatar que a conjugação da informação proveniente das categorias e itens conduziu a estes resultados.

> *Observação 4*: A implementação da abordagem em que se utilizam os 5 melhores vizinhos obtidos a partir do conjunto de dados dos itens e das categorias (*Cit5*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas o conjunto de treino das categorias (*Cat*).

Ao comparar os resultados obtidos nas duas abordagens verifica-se que *Cit5* é melhor em todos os parâmetros e métricas com excepção do parâmetro "10 itens" nas métricas *Recall* e *F1*. Relativamente à percentagem de sessões com previsão observa-se uma diferença significativa, de tal forma que *Cit5* apresenta uma média de previsão de 71,4% enquanto *Cat* de 14,7%.

Como esperado verifica-se que *Cit5* globalmente apresenta melhores resultados que *Cat* permitindo constatar que tal como na *Observação 3* a conjunção dos conjuntos de dados das categorias e itens possibilitou a obtenção de melhores resultados.

Ao comparar os resultados obtidos em *Cit5* e *Cit* observa-se que *Cit5* é melhor na métrica *Precision* em todos os parâmetros com a excepção de "All but 1", em *Recall* só é melhor em "5 itens" e em *F1* é melhor nos parâmetros "2 itens" e "5 itens". É de salientar o resultado obtido em "All but 1" pois indicia que a abordagem *Cit* pode ter melhores resultados quando existe um histórico de maior dimensão e mais informação disponível acerca dos itens visitados pelo utilizador. Ao nível da percentagem de sessões com previsão observa-se uma diferença significativa com pendor para *Cit5* com uma média de 71,4% enquanto que *Cit* apresenta apenas 27,3%.

Tal como para as subcategorias também se verifica uma relação clara entre a utilização do conjunto de dados relativo aos itens e o aumento da percentagem das sessões com previsão.

A tabela 4.6 visa mostrar os resultados mais elevados obtidos nesta secção para cada métrica e parâmetro, identificando também qual a abordagem e percentagem de sessões com previsão inerente. Idêntica análise se pode fazer através dos gráficos apresentados em 4.4 e 4.5. Das abordagens que dão origem aos resultados apresentados nesta tabela verifica-se uma presença significativa da abordagem *Sub*, pois é melhor em 5 dos cenários, sendo melhor para todas as métricas no parâmetro "10 itens" e em "2 itens" com excepção do *Recall*. A abordagem *Cit* apresenta um bom desempenho em "All but 1" e *Cit5* em "5 itens". É de salientar que *Sit* não dispõe de nenhum valor nesta tabela e que as abordagens que utilizam o conjunto de dados das subcategorias predominam, uma vez que estão presentes em 7 dos 12 cenários.

Tendo em conta a hipótese colocada no início desta secção e toda a análise apresentada, constata-se que tal como esperado as abordagens propostas que integraram a informação relativa às subcategorias ou categorias e aos itens, conseguiram globalmente obter melhores resultados, embora a abordagem que utiliza apenas a informação relativa às subcategorias, ao contrário do esperado, tenha melhores resultados ao nível dos parâmetros "2 itens" nas métricas *Precision* e *F1* e "10 itens" em todas as métricas.

	Dados 2 itens	**Dados 5 itens**	**Dados 10 itens**	**"All but 1"**
AVG(precision)(%)	14,3	8,8	4,8	5,7
Sessões com Prev. (%)	24,9	71,2	29,2	70,6
Abordagem	*Sub*	*Cit5*	*Sub*	*Sit5*
AVG(recall)(%)	22,9	9,5	7,1	44,1
Sessões com Prev. (%)	24,1	75,8	29,2	29,0
Abordagem	*Cit*	*Sit5*	*Sub*	*Cit*
AVG(f1)(%)	16,3	8,9	5,7	9,9
Sessões com Prev. (%)	24,9	71,2	29,2	29,0
Abordagem	*Sub*	*Cit5*	*Sub*	*Cit*

Tabela 4.6: Melhores resultados obtidos na secção 4.3 e respectiva tabela onde foi apresentado o resultado.

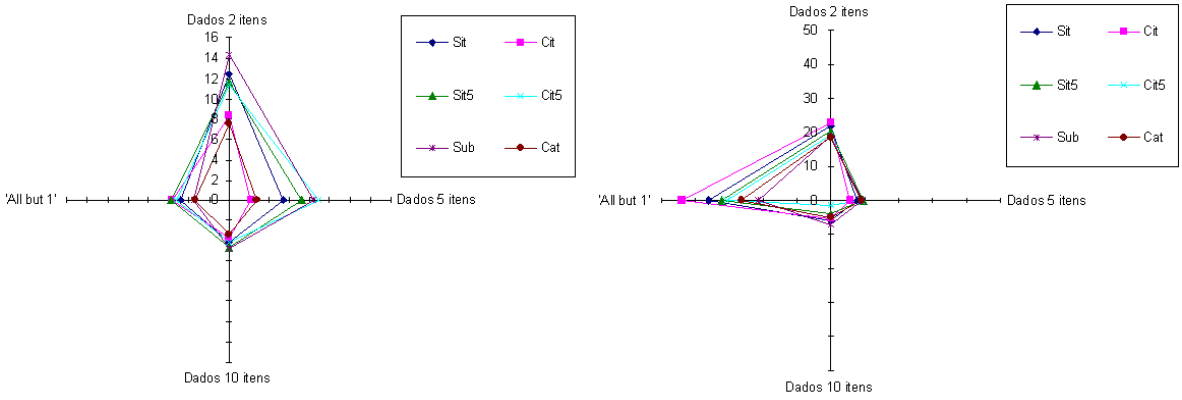


Figura 4.4: Comparação gráfica das várias abordagens e parâmetros para as medidas de avaliação *Precision* (gráfico à esquerda); e *Recall*.

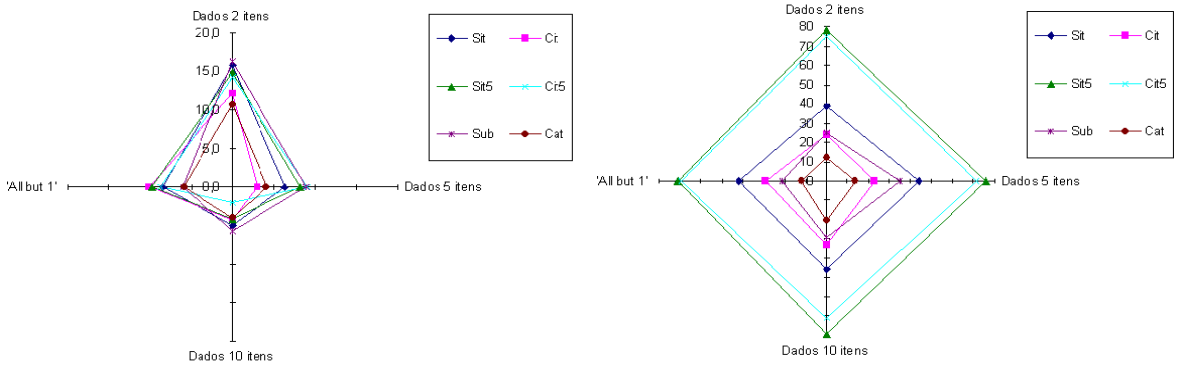


Figura 4.5: Comparação gráfica das várias abordagens e parâmetros para a medida de avaliação *F1* (gráfico à esquerda); e percentagem de sessões com previsão.

4.4 Integração de informação relativa ao tempo

Esta secção visa aplicar o método descrito em 3.6.2, com o propósito de integrar a informação relativa ao intervalo de tempo de visualização. Pretende-se assim utilizar o conjunto de dados dos itens em conjunto com a informação do tempo de visualização sem utilidade ou com utilidade, tendo em conta as duas funções propostas em 3.6.2, com o objectivo de verificar qual deles consegue alcançar melhores resultados.

A utilização desta informação tem como base uma das abordagens implementada na secção 4.2, mais concretamente a última abordagem caracterizada pela Utilização da Correlação, *Inverse User Frequency* e *Default Vote* para o utilizador activo e utilizador a comparar (*Cout*).

Como primeira abordagem procedeu-se à utilização da informação relativa aos itens e ao intervalo de tempo de visualização, para obter previsões dos itens. Nesta situação os vizinhos são calculados através da soma dos pesos obtidos a partir dos votos nos itens e a partir do tempo (wtp + w). Os resultados obtidos são apresentados na coluna "*Tp*" da tabela 4.7.

Para além da abordagem apresentada anteriormente, procedeu-se à utilização de uma abordagem semelhante, onde se propõe a utilização da função $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$ para calcular a utilidade do tempo de visualização. Neste caso os vizinhos são calculados através da soma dos pesos obtidos a partir dos votos nos itens e da utilidade do tempo ($wut + w$). Os resultados obtidos são apresentados na coluna "*UtFlg*" da tabela 4.7.

Outra abordagem utilizada é semelhante à anterior, com a excepção da função de utilidade que é $ln(x/6) + 4 - (sqrt(x/3))$. Os resultados obtidos são apresentados na coluna "*UtFln*" da tabela 4.7.

Em seguida vai realizar-se a análise de resultados com intuito de verificar se é valida a hipótese formulada no inicio desta secção. Esta hipótese foi subdividida em duas observações principais mais especificas com o intuito de facilitar a análise.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)			Sessões com Prev. (%)		
	Tp	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*
2 itens	7,1	7,2	7,1	26,7	26,9	26,7	11,2	11,4	11,2	93,7	93,7	93,7
5 itens	7,3	7,3	6,5	15,3	15,3	14,8	9,9	9,9	9,0	93,9	93,9	93,9
10 itens	3,3	3,3	3,3	8,2	8,2	8,2	4,7	4,7	4,7	100,0	100,0	100,0
"All but 1"	4,0	4,0	4,0	40,8	40,8	40,9	7,3	7,3	7,3	90,4	90,4	90,4

Tabela 4.7: Resultados obtidos a partir da integração de informação relativa ao intervalo de tempo de visualização.

> *Observação 1*: A implementação da abordagem em que se utilizam o conjunto de treino dos itens e a função de utilidade $3+(lg(x)-0.25/sqrt(x))-sqrt(x/3)$ (*UtFlg*) apresenta melhores resultados relativamente à abordagem em que se utiliza o conjunto de treino dos itens e o intervalo de tempo de visualização sem utilizar função de utilidade (*Tp*).

Comparando os resultados obtidos a partir das duas abordagens verifica-se uma grande semelhança. As únicas diferenças podem observar-se ao nível do parâmetro "2 itens" para todas as métricas, onde *UtFlg* apresenta melhores resultados, embora com diferenças pouco significativas. Ao nível da percentagem de sessões com previsão constata-se que são iguais com uma média de 94,5%.

Contrariamente ao esperado não se verificaram diferenças significativas entre os resultados obtidos pelas duas abordagens.

Observação 2: A implementação da abordagem em que se utilizam o conjunto de treino dos itens e a função de utilidade $ln(x/6) + 4 - (sqrt(x/3))$ (*UtFln*) apresenta melhores resultados relativamente à abordagem em que se utiliza o conjunto de treino dos itens e o intervalo de tempo de visualização sem utilizar função de utilidade (*Tp*).

Comparando os resultados obtidos a partir das duas abordagens verifica-se uma grande semelhança. As únicas diferenças podem observar-se ao nível do parâmetro "5 itens" para todos os parâmetros, onde *Tp* apresenta melhores resultados, embora com diferenças pouco significativas. A abordagem *UtFln* só é maior na métrica *Recall* e parâmetro "All but 1". Ao nível da percentagem de sessões com previsão constata-se são iguais com uma média de 94,5%.

Contrariamente ao esperado não se verificaram diferenças significativas entre os resultados obtidos pelas duas abordagens.

Ao comparar *UtFlg* com *UtFln* verifica-se que *UtFlg* apresenta melhor resultado nos parâmetros "Dados 2 itens" e "Dados 5 itens" em todas as métricas. Verifica-se um resultado inferior para o parâmetro "All but 1" na métrica *Recall* e resultados iguais para os restantes valores. O número de sessões com previsão também é igual.

A partir dos gráficos 4.6 e 4.7 pretendem-se representar gráficamente os resultados obtidos nesta secção para cada métrica e parâmetro. Perante as hipóteses apresentadas nesta secção constatou-se que os resultados obtidos não estão de acordo com o esperado, pelo que da utilização das funções utilidade não se constatou grande mais-valia, pelo menos nesta fase. Relativamente à comparação das duas funções utilidade verificou-se que *UtFlg* apresenta melhores resultados, embora com diferenças muito pequenas. Pelo facto das diferenças serem pouco significativas não se pode afirmar que a função *UtFlg* modela melhor o comportamento dos utilizadores ao nível das suas preferências que *UtFln*.

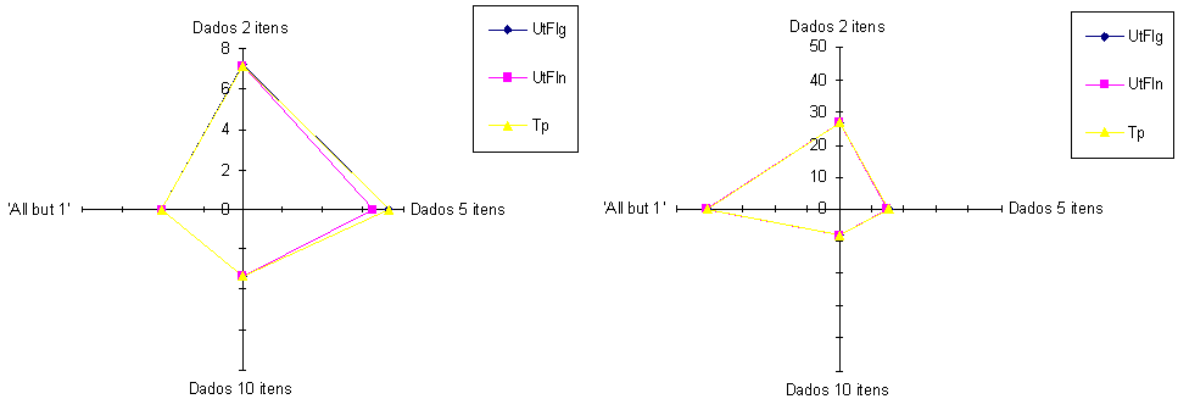


Figura 4.6: Comparação gráfica das várias abordagens e parâmetros para as medidas de avaliação *Precision* (gráfico à esquerda); e *Recall*.

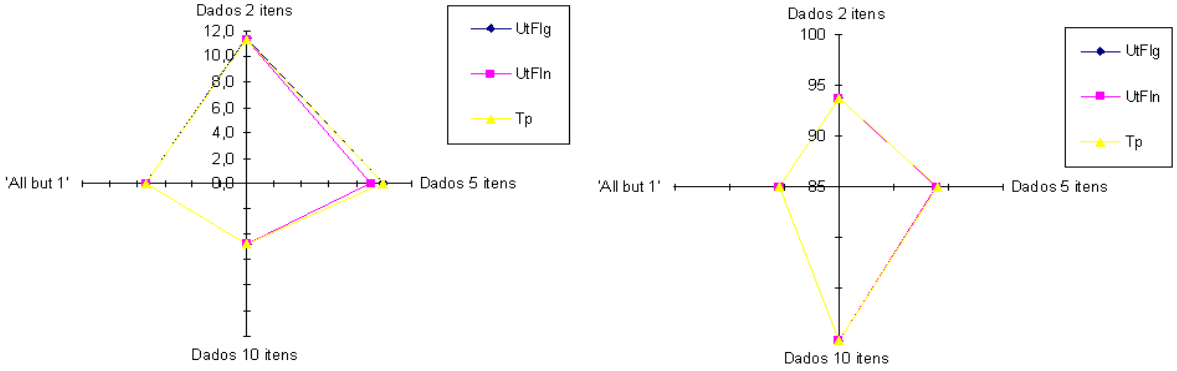


Figura 4.7: Comparação gráfica das várias abordagens e parâmetros para a medida de avaliação *F1* (gráfico à esquerda); e percentagem de sessões com previsão.

4.5 Comparação e análise relativa às secções anteriores

Esta secção tem como propósito a comparação dos melhores resultados obtidos para cada uma das abordagens analisadas nas secções anteriores. Para realizar esta comparação procedeu-se à recolha dos melhores valores obtidos em cada uma das secções anteriores, que analisaram as diferentes abordagens para a obtenção de recomendações. Esta informação foi compilada numa tabela 4.8, com informação relativa à medida, parâmetro, percentagem de sessões com previsão e a referência para a abordagem que deu origem ao respectivo valor.

	Dados 2 itens	**Dados 5 itens**	**Dados 10 itens**	**"All but 1"**
AVG(precision)(%)	14,3	8,8	5,3	5,7
Sessões com Prev. (%)	24,9	71,2	25,7	70,6
Abordagem	*Sub*	*Cit5*	*Cout*	*Sit5*
AVG(recall)(%)	39,1	30,7	25,7	49,2
Sessões com Prev. (%)	98,4	97	100	93,8
Abordagem	*Corr*	*Corr*	*Corr*	*Corr*
AVG(f1)(%)	16,3	9,9	6,3	9,9
Sessões com Prev. (%)	24,9	93,9	100	29
Abordagem	*Sub*	*Tp*; *UtFlg*	*Corr*	*Cit*

Tabela 4.8: Melhores resultados obtidos e respectiva abordagem onde foi apresentado o resultado.

Quando conhecidos 2 itens, verifica-se que com a utilização do conjunto de dados das subcategorias (*Sub*) se obtêm melhores valores para as medidas *Precision* e *F1*. Para a medida *Recall*, o melhor resultado é obtido através da utilização da correlação isoladamente (*Corr*).

No que diz respeito ao parâmetro experimental "Dados 5 itens", constata-se que os melhores resultados estão distribuídos por três abordagens, ao nível da medida *Precision* o melhor resultado é obtido quando utilizado o conjunto de treino das categorias, ao nível da medida *Recall* o melhor resultados verifica-se na utilização simples da correlação. Ao nível da medida *F1* o melhor resultado é observado quando é utilizado o tempo, com e sem utilidade.

Quando conhecidos 10 itens observa-se que: para a medida *Precision* o melhor resultado foi obtido através da utilização da abordagem *Cout*, onde é utilizada a abordagem da Correlação, *Inverse User Frequency* e *Default Vote* para o utilizador activo e utilizador a comparar. As medidas *Recall* e *F1* têm o melhor resultado na abordagem da correlação.

Relativamente ao parâmetro experimental "All but 1" verifica-se que para a medida *Precision* o melhor valor resulta da utilização das subcategorias. Ao nível do *Recall* o melhor valor é obtido pela correlação simples. Ao nível da medida *F1* o melhor resultado é obtido através da abordagem *Cit*, que utiliza dados relativos às categorias.

No que diz respeito à percentagem de sessões com previsão verifica-se que a média é maior na abordagem que utiliza a informação relativa ao tempo (95%) e na abordagem que utiliza o conjunto de dados dos itens (95%) e menor quando utilizada informação relativa às categorias e subcategorias (43%).

Quanto à média da percentagem de sessões com previsão por parâmetro, constata-se que a parâmetro experimental com maior percentagem é "Dados 10 itens" com 73%, seguido de "Dados 5 itens" e "Dados 2 itens" com 69%, enquanto que "All but 1" tem 66%.

Do ponto de vista das métricas de avaliação pode observar-se que a *Precision* engloba três resultados obtidos a partir da utilização do conjunto de dados das categorias e subcategorias e apenas um proveniente da utilização do conjunto de dados dos itens. Ao nível do *Recall* a correlação domina em todos os parâmetros. A métrica *F1* é a que apresenta maior diversidade pois apresenta resultados das três abordagens utilizadas nas secções anteriores.

Ao analisar a tabela como um todo pode contactar-se que na sua maioria os melhores resultados provêm da secção onde se utiliza apenas o conjunto de treino dos itens (4.2), seguida da secção onde se utiliza o conjunto de treino das categorias e subcategorias (4.3).

Outra análise que se pode realizar está relacionada com as abordagens que apresentam maior percentagem de sessões com previsão por parâmetro. A tabela 4.9 apresenta os maiores valores para as percentagens de sessões com previsão, tendo em conta cada um dos parâmetros. Após análise, verifica-se que há uma abordagem com resultados em todos os parâmetros, que corresponde à abordagem que utiliza o algoritmo da Correlação isoladamente. Constata-se também a inexistência de referencias à abordagem que utiliza informação relativa às categorias e subcategorias.

Parâmetro	**Sessões com Previsão (%)**	**Tabelas**
Dados 2 itens	98,4	*Corr*
Dados 5 itens	97,0	*Corr*; *CIF*; *Cact*
Dados 10 itens	100	*Corr*; *CIF*; *Cact*; *Tp*; *UtFlg*; *UtFln*
'All but 1'	93,8	*Corr*

Tabela 4.9: Maiores valores de percentagem de sessões com previsão por parâmetro.

4.6 Escolha dos melhores itens para previsão

Nesta fase de experimental pretende-se sugerir um número limitado de itens, escolhidos tendo em conta o seu valor previsto. Na fase anterior todos os itens previstos eram utilizados para avaliar as abordagens, enquanto que nesta fase são apenas sugeridos um, três, cinco e dez itens com os maiores valores previstos. Só depois de escolhidos os valores mais elevados é que são aplicadas as medidas de avaliação *Precision*, *Recall* e *F1*.

Como método para obtenção dos melhores vizinhos e respectivos itens associados, foi utilizado o método já mencionado anteriormente caracterizado pela utilização da Correlação, *Default Vote* e *Inverse User Frequency* e o parâmetro dados 2 itens.

Nesta secção pretendem-se comparar os resultados obtidos utilizando apenas informação relativa aos itens (votos) e os resultados obtidos utilizando a informação dos itens em conjunto com a informação relativa ao intervalo de tempo de visualização ou utilizando apenas informação relativa ao tempo.

Na primeira abordagem utilizou-se apenas a informação relativa aos itens para calcular os melhores itens. Os resultados obtidos são apresentados na tabela 4.10 (*Itens*).

A segunda abordagem tem em conta o caso em que $v_{i,j}$ é igual $\overline{v}_i$, o que causa a anulação de um dos factores. Se esta condição se verificar para todos os itens então $p_{a,j} = \overline{v}_a$ (ver equação 2.1), ocorrendo assim perda de informação. Posto isto, propõe-se a utilização da informação relativa ao tempo quando esta condição se verifica (ver 3.6.2). A utilização do tempo é realizada com e sem utilidade sendo identificados na tabela 4.11 pelas seguintes abreviaturas, sem utilidade *Tp*, com utilidade *UtFlg* quando $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$ e *UtFln* quando $ln(x/6) + 4 - (sqrt(x/3))$.

Na terceira abordagem pretende-se verificar qual a importância e qual o impacto de utilizar apenas a informação relativa ao tempo para calcular os melhores itens. Para levar a cabo esta operação substituiu-se o factor $v_{i,j} - \overline{v}_i$ pelo resultado do tempo (enquanto que no método anterior só acontecia quando este factor era 0). A utilização do tempo é realizada com e sem utilidade sendo identificados na tabela 4.12 pelas seguintes abreviaturas, sem utilidade *Tp1*, com utilidade *UtFlg1* quando $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$ e *UtFln1* quando $ln(x/6) + 4 - (sqrt(x/3))$.

Na quarta abordagem devido aos bons resultados obtidos com a terceira abordagem procedeu-se à utilização do tempo independentemente de $v_{i,j} - \overline{v}_i$ ser diferente de 0 (na segunda abordagem considerou-se quando igual a 0). Para realizar tal tarefa procedeu-se à soma do valor do tempo, da seguinte forma $w(a,i) * ((v_{i,j} - \overline{v}_i) + tp_{i,j})$. A utilização do tempo é realizada com e sem utilidade sendo identificados na tabela 4.13 pelas seguintes abreviaturas, sem utilidade *Tp2*, com utilidade *UtFlg2* quando $3 + (lg(x) - 0.25/sqrt(x)) - sqrt(x/3)$ e *UtFln2* quando $ln(x/6) + 4 - (sqrt(x/3))$.

Parâmetro	**AVG(precision)(%)**	**AVG(recall)(%)**	**AVG(f1)(%)**
1 item	8,4	2,5	3,9
3 itens	8,9	7,4	8,1
5 itens	9,5	11,9	10,6
10 itens	9,7	18,1	12,6

Tabela 4.10: Utilização da Correlação, *Inverse User Frequency* e *Default Vote* para o utilizador activo e utilizador a comparar tendo em conta os itens com maior previsão.

Em seguida vai realizar-se a análise de resultados com intuito de verificar se é valida a hipótese formulada no inicio desta secção. Esta hipótese foi subdividida em três observações principais mais especificas com o intuito de facilitar a análise e acomodar melhor as abordagens apresentadas anteriormente.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)		
	Tp	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*
1 item	10,1	11,0	10,1	2,5	2,8	2,5	4,0	4,4	4,0
3 itens	10,5	10,2	10,4	8,1	7,9	7,9	9,1	8,9	8,9
5 itens	10,2	10,1	10,3	13,1	12,8	13,3	11,5	11,3	11,6
10 itens	9,9	9,9	9,8	18,3	18,3	18,2	12,8	12,8	12,8

Tabela 4.11: Resultados obtidos a partir do valor do intervalo de tempo de visualização e os votos para o calculo dos melhores itens.

Observação 1: A implementação da abordagem em que se utilizam os itens e a informação relativa ao tempo caso $v_{i,j}$ igual $\overline{v}_i$ (*Tp*, *UtFlg* e *UtFln*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas os itens (*Itens*).

Comparando os resultados obtidos pelas abordagens verifica-se que todos os parâmetros, métricas e abordagens apresentam resultados maiores ou iguais que a abordagem *Itens*.

Tal como esperado verificou-se que a integração da informação relativa ao tempo caso $v_{i,j}$ igual $\overline{v}_i$ conduziu à obtenção de melhores resultados.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)		
	Tp1	*UtFlg1*	*UtFln1*	*Tp1*	*UtFlg1*	*UtFln1*	*Tp1*	*UtFlg1*	*UtFln1*
1 item	10,6	11,5	9,7	2,8	2,9	2,5	4,5	4,6	3,9
3 itens	11,4	11,5	11,4	8,9	9,0	8,8	10,0	10,1	9,9
5 itens	10,8	10,8	10,7	14,1	14,1	13,9	12,2	12,2	12,1
10 itens	10,2	10,2	10,2	19,9	19,7	19,7	13,5	13,4	13,4

Tabela 4.12: Resultados obtidos a partir da utilização apenas do valor do intervalo de tempo de visualização para o calculo dos melhores itens.

Observação 2: A implementação da abordagem em que se utiliza apenas informação relativa ao tempo (*Tp1*, *UtFlg1* e *UtFln1*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas os itens (*Itens*).

Comparando os resultados obtidos pelas abordagens verifica-se que todos os parâmetros, métricas e abordagens apresentam resultados maiores que a abordagem *Itens*. Com a excepção de *UtFln1* em "1 item" nas métricas *Recall* e *F1* que são iguais.

Tal como esperado verificou-se que a utilização apenas da informação relativa ao tempo permitiu a obtenção de melhores resultados.

Parâmetros	AVG(precision)(%)			AVG(recall)(%)			AVG(f1)(%)		
	Tp2	*UtFlg2*	*UtFln2*	*Tp2*	*UtFlg2*	*UtFln2*	*Tp2*	*UtFlg2*	*UtFln2*
1 item	11,5	11,5	11,9	2,9	2,5	2,9	4,6	4,1	4,7
3 itens	11,1	10,8	11,1	8,6	8,3	8,6	9,7	9,4	9,7
5 itens	10,7	10,6	11,0	13,7	13,8	14,4	12,0	12,0	12,4
10 itens	10,0	10,2	10,2	18,8	19,5	19,5	13,0	13,4	13,3

Tabela 4.13: Resultados obtidos a partir da soma do valor do intervalo de tempo de visualização e os votos para o calculo dos melhores itens.

> *Observação 3*: A implementação da abordagem em que se utilizam os itens e a informação relativa ao tempo (*Tp2*, *UtFlg2* e *UtFln2*) apresenta melhores resultados relativamente à abordagem em que se utiliza apenas os itens (*Itens*).

Comparando os resultados obtidos pelas abordagens verifica-se que todos os parâmetros, métricas e abordagens apresentam resultados maiores que a abordagem *Itens*. Com a excepção de *UtFlg2* em "1 item" na métrica *Recall* que é igual.

Tal como esperado verificou-se que o somatório da informação relativa ao tempo e da informação relativa aos itens permitiu a obtenção de melhores resultados.

A tabela 4.14 visa compilar os melhores resultados e respectivas abordagens. Através desta pode observar-se a prevalência de uma abordagem de recomendação para os diferentes parâmetros, ou seja, para o parâmetro de recomendação de "1 item", a existência de resultados provenientes da abordagem *UtFln2*, o mesmo se pode afirmar para a recomendação de "3 itens" e *UtFlg2*, assim como para a recomendação de "5 itens" onde a abordagem que prevalece é a *UtFln2*. Na recomendação de "10 itens" também se verifica a existência de denominador comum que é a abordagem *Tp1*.

	1 item	**3 itens**	**5 itens**	**10 itens**
AVG(precision)(%)	11,9	11,5	11,0	10,2
Abordagem	*UtFln2*	*UtFlg1*	*UtFln2*	*Tp1*; *UtFlg1*; *UtFln1*; *UtFlg2*; *UtFln2*
AVG(recall)(%)	2,9	9,0	14,4	19,9
Abordagem	*UtFlg1*;*Tp2*;*UtFln2*	*UtFlg1*	*UtFln2*	*Tp1*
AVG(f1)(%)	4,7	10,1	12,4	13,5
Abordagem	*UtFln2*	*UtFlg1*	*UtFln2*	*Tp1*

Tabela 4.14: Melhores resultados obtidos para a recomendação de itens.

Relativamente à análise das abordagens com melhores resultados por métrica constata-se maior diversidade, embora se possa salientar a presença em todas as métricas das abordagens *UtFln2*, *UtFlg1* e *Tp1*.

Pode-se verificar ainda que a abordagem *UtFln2* é a mais referenciada, em 7 vezes, seguida de *UtFlg1* com 5 referencias e *Tp1* com 3 referências. Nesta análise incluem-se as referências múltiplas.

Importa ainda realizar outra análise relacionada com o tempo. A tabela 4.15 tem como intuito estabelecer a relação entre os parâmetros, as métricas de avaliação e o tempo (*Tp*) e sua utilidade (*UtFlg* e *UtFln*); identificando o número de vezes que cada uma delas apresentou melhores resultados relativamente à outra. O tempo foi utilizado em três abordagens, como tal 100% corresponde há ocorrencia

de melhores resultados de uma utilização do tempo relativamente às outras em todas as abordagens apresentadas nesta secção. Em caso de igualdade foi considerada a divisão da percentagem pelo número de abordagens com melhores resultados. Portanto caso as três *Tp*, *UtFlg* e *UtFln* apresentem o mesmo valor a percentagem atribuída a cada é de 11,1 $(33,4/3)$.

	Precision			***Recall***			***F1***		
Abordagem	*Tp*	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*	*Tp*	*UtFlg*	*UtFln*
1 item	33,4%	33,4%	33,4%	16,7%	66,8%	16,7%	0%	66,8%	33,4%
3 itens	50,1%	33,4%	16,7%	50,1%	33,4%	16,7%	50,1%	33,4%	16,7%
5 itens	16,7%	16,7%	66,8%	16,7%	50,1%	33,4%	16,7%	16,7%	66,8%
10 itens	22,2%	33,3%	22,2%	50,1%	33,4%	16,7%	44,5%	44,5%	11,1%

Tabela 4.15: Relação entre os parâmetros, as métricas de avaliação e as abordagens de utilização da informação relativa ao tempo

Através da observação da tabela consegue-se inferir que quando recomendado "1 item" a *UtFlg* apresenta maior percentagem de melhores resultados nas métricas *Recall* e *F1*. Na recomendação de "3 itens" a utilização do tempo isoladamente (*Tp*) apresenta-se como a melhor alternativa em todas as métricas. Quando recomendados "5 itens" a alternativa com melhores resultados é *UtFln* para as métricas *Precision* e *F1*. Ao nível da recomendação de "10 itens" as alternativas *Tp* e *UtFlg* apresentam os melhores resultados.

Outra análise que se pode executar está relacionada com o ganho associado ao facto de recomendar 1, 3, 5, ou 10 itens. A tabela 4.16 apresenta o valor absoluto da diferença dos valores de *F1* obtidos em função do número de itens recomendados (retirados da tabela 4.14).

Diferença entre parâmetros	**AVG(f1)(%)**
1 item - 3 itens	5,4
3 itens - 5 itens	2,3
5 itens - 10 itens	1,1

Tabela 4.16: Valor absoluto da diferença obtida entre resultados obtidos dependendo do número de itens a recomendar.

Como se pode observar o ganho vai diminuindo com o aumento do número de itens a recomendar. Isto é, o ganho de recomendar "3 itens" em vez de "1 item" é superior ao ganho de recomendar "5 itens" em vez de 3 itens e de recomendar 10 em vez de 5.

Capítulo 5

Conclusões

Nos dias que correm o tempo é um bem essencial, e como tal, nenhum cliente ou comprador quer perder muito tempo na procura de determinado produto, cujas características se adequam convenientemente às suas necessidades. No passado foram aplicadas determinadas técnicas que visavam responder a esta exigência ao nível das lojas físicas, conduzindo a um maior interesse com esta temática. Actualmente, a maior facilidade de acesso à `Internet` impulsionou e proporcionou um crescimento da sua utilização, que por sua vez acarretou uma diversificação da sua utilização, de tal forma que o número de transacções realizadas através da `Internet`, apresenta um crescimento considerável. Estando cientes desta realidade, muitas empresas encararam esta oportunidade e lançaram-se no desafio da `Internet`, criando lojas *on-line*. Devido às características destas lojas as empresas foram colocadas perante um problema (que já tinham tratado para as lojas físicas), relacionado com o acesso aos produtos, de forma a proporcionar maior satisfação e rapidez na realização das compras. Como resposta a este problema começaram a surgir vários estudos que recuperam algumas técnicas provenientes do domínio da "Information retrieval", "Data mining", estatística, etc., e assim nasceram os `Sistemas de Recomendação`.

Da implementação e estudo dos `Sistemas de Recomendação` surgiram essencialmente três grandes abordagens para encarar o problema: baseada no conteúdo, *Collaborative Filtering* e abordagem híbrida, cada uma delas com as suas variações, vantagens e desvantagens (ver capítulo 2). A abordagem utilizada neste estudo foi o *Collaborative Filtering*. Os algoritmos e variantes implementados foram: Correlação, *Default Vote* e *Inverse User Frequency*. Procurou-se também realizar a conjugação da utilização dos algoritmos e variantes, com o intuito de retirar maior partido das suas particularidades e da sua especificidade. Para além desta metodologia, tentou-se ainda enriquecer a informação inicial, que apenas considerava os acessos presentes nos *Web Access Logs* aos itens, acrescentado informação relativa às categorias e subcategorias dos itens. Outra informação utilizada, foi o intervalo de tempo que determinado utilizador permanece na página de um item, obtido a partir da diferença do tempo de acesso às páginas. Variada literatura consultada durante a pesquisa bibliográfica remete para a relevância existente na informação relativa ao intervalo de tempo de visualização, pois está relacionada com o interesse por determinado item.

Os três parágrafos que se seguem pretendem apresentar as conclusões retiradas em cada uma das secções que abordaram a escolha dos melhores vizinhos. As conclusões relativas à escolha dos itens com maior valor previsto para realizar a recomendação, são apresentadas mais à frente neste capítulo.

A hipótese considerada na primeira secção tinha como propósito comparar os resultados do algoritmo de correlação isoladamente e em conjunto com as suas variantes *Inverse User Frequency* e *Default Vote* com o intuito de verificar qual deles consegue alcançar melhores resultados. Tendo em

conta a métrica *F1* pois é a que permite a avaliação completa do problema, pode concluir-se que a hipótese foi comprovada em três dos quatro parâmetros experimentais, a única excepção obteve-se quando conhecidos "10 itens" (tabela 4.3).

A hipótese considerada na segunda secção tinha como propósito comparar a utilização do conjunto de dados das subcategorias ou categorias isoladamente e em conjunto com a informação relativa aos itens com o propósito de verificar qual deles consegue alcançar melhores resultados. Tendo em conta a métrica *F1* pode concluir-se que a hipótese colocada se mostrou válida em dois ("5 itens" e "All but 1") dos quatro parâmetros considerados (tabela 4.6).

Por fim a hipótese considerada na terceira secção tinha como propósito comparar a utilização do conjunto de dados dos itens e da informação do intervalo de tempo de visualização sem utilidade, com a utilização das duas funções de utilidade propostas, com o objectivo de verificar qual delas consegue alcançar melhores resultados. Tendo em conta a métrica *F1* pode concluir-se que a hipótese colocada se mostrou válida em apenas um ("2 itens") dos quatro parâmetros considerados (tabela 4.7). Outra questão importante nesta secção era avaliar quais das funções de utilidade modela melhor o comportamento dos utilizadores ao nível das suas preferências. Posto isto, devido ao facto de se verificarem poucas diferenças de resultados e as que existem serem pouco significativas nada se pode concluir relativamente a esta questão.

Da comparação e análise dos melhores resultados obtidos na escolha dos melhores vizinhos constatou-se que, embora se verifique uma diferença pouco significativa entre os resultados gerados a partir das diversas abordagens implementadas, pode concluir-se que: há uma tendência para a variabilidade nos métodos que apresentam os melhores resultados no que concerne à escolha dos melhores vizinhos do utilizador activo, pelo que, numa eventual utilização da implementação apresentada nesta dissertação, num sítio de comércio electrónico, o `Sistema de Recomendação` deveria consistir num processo dinâmico de escolha da abordagem utilizada para obter as recomendações, tendo em conta os itens conhecidos, isto é, os itens já vistos pelo utilizador. Por exemplo no caso do utilizador activo já ter visualizado 2 itens o SR deveria utilizar a abordagem *Sub* caracterizada pela utilização do conjunto das subcategorias; caso o utilizador activo tivesse visto 10 itens deveria utilizar-se a abordagem *Corr* que utiliza a correlação isoladamente e o conjunto de dados dos itens (tabela 4.8).

Também ao nível do número de sessões com previsão, têm que se tomar decisões, na altura da aplicação da implementação num sítio de comércio electrónico, pois dependendo dos itens conhecidos podem obter-se maior ou menor número de previsões. Em alguns casos o "responsável" do sítio pode ponderar a utilização de abordagens com maior número de sessões com pouco prejuízo da métrica. Por exemplo no caso do utilizador activo já ter visualizado 2 itens a abordagem com melhor resultado é *Sub*, no entanto realiza previsões em apenas 24,9% das sessões, portanto o "responsável" do sítio pode achar interessante a utilização do segundo melhor *UtFlg* que apresenta uma percentagem de previsão bastante superior, de 93,7%, no entanto terá que assumir a diminuição da métrica de 4,9%.

Ao comparar as abordagens com maior percentagem de sessões com previsão (tabela 4.9) e as abordagens detentoras dos melhores resultados (tabela 4.8), constata-se que apenas o resultado de um parâmetro é coincidente nesta duas tabelas. Conclui-se assim que o maior número de sessões com previsão não implica melhores resultados, dai que a escolha deste critério tenha de ser bastante ponderada. Pelo que o critério a considerar deve ser preferencialmente a métrica *F1*.

Relativamente à temática da escolha dos melhores vizinhos, pode-se ainda inferir que todas as abordagens consideradas contribuíram, em diferentes cenários de utilização (diferentes parâmetros e medidas) para a obtenção de resultados considerados como os melhores. Isto é, ao nível da métrica *F1* quando conhecidos "2 itens" a abordagem que melhor resultado obteve, tem origem a partir da da utilização do conjunto de dados das subcategorias, enquanto que quando conhecidos "5 itens" as abordagens com melhores resultados provêem da utilização do tempo; quando conhecidos "10 itens"

a abordagem com melhor resultado utiliza apenas o conjunto de dados dos itens e por fim quando utilizado o "All but 1" a abordagem com melhores resultados utiliza o conjunto de dados das categorias.

No que diz respeito à escolha dos itens com maior valor previsto para realizar a recomendação, onde se colocou a hipótese de comparar os resultados obtidos utilizando apenas informação relativa aos itens (votos) e os resultados obtidos utilizando a informação dos itens em conjunto com a informação relativa ao intervalo de tempo de visualização ou utilizando apenas informação relativa ao tempo, concluiu-se, tendo em conta a métrica *F1* que a hipótese colocada se mostrou válida em todos os parâmetros considerados (tabela 4.14). Pode concluir-se assim que a incorporação da informação relativa ao tempo conduziu a uma melhoria dos resultados. Também se pode concluir que a utilização da informação relativa ao tempo apresenta melhores resultados quando utilizada independentemente de $(v_{i,j} - \overline{v}_i)$ ser nulo ou não. Isto é, quando o factor mencionado é nulo utiliza-se só o tempo, quando não é nulo adiciona-se a informação relativa ao tempo à informação dos votos.

Tendo em conta esta fase da recomendação (escolha dos itens com melhores previsões), verifica-se que existe uma certa tendência para obtenção de melhores resultados, em diferentes abordagens, para diferentes parâmetros, pelo que antes da implementação do `Sistema de Recomendação` terá que se decidir qual o número de itens a recomendar e consequentemente qual a abordagem a utilizar. Por exemplo se o "responsável" pelo sítio pretender recomendar apenas um item, devido a preocupações relativas ao espaço disponível na página do sítio, deve utilizar a abordagem *UtFln2* (tabela 4.14).

Por outro lado também se deve considerar qual o ganho associado ao facto de recomendar n itens. Como se pôde constatar na análise de resultados o ganho associado ao facto de se recomendar "3 itens" em detrimento de 1 é maior que nas restantes alternativas (tabela 4.16). Assim o "responsável" pelo sítio pode decidir recomendar "3 itens", em vez de 1 pois o ganho associado é significativo, no entanto o mesmo já não se sucede nos restantes cenários.

Relativamente às abordagens relacionadas ao tempo (tabela 4.15), utilizadas na secção dos melhores itens, tendo em conta a métrica *F1*, pode concluir-se que há uma tendência para a diversidade de abordagens com melhores resultados, pois quando recomendado "1 item" a abordagem que utiliza a função de utilidade $3+(lg(x)-0.25/sqrt(x))-sqrt(x/3)$ é a que apresenta melhores resultados, para a recomendação de "2 itens" surge a utilização do tempo sem utilidade, para "5 itens" temos a função de utilidade $ln(x/6)+4-(sqrt(x/3))$ e na recomendação de "10 itens" surge uma igualdade entre a utilização do tempo sem utilidade e a utilização da função de utilidade $3+(lg(x)-0.25/sqrt(x))-sqrt(x/3)$.

5.1 Trabalho Futuro

Como trabalho futuro seria interessante implementar o `Sistema de Recomendação` estudado nesta dissertação e disponibilizar recomendações no sítio `WEB`. Adicionalmente seria interessante recolher a opinião e avaliação dos utilizadores relativamente à qualidade (interesse e utilidade) das recomendações. Com a avaliação realizada pelos utilizadores poderia-se inferir acerca da capacidade do `Sistema de Recomendação` possibilitar a diminuição do tempo de procura e aumentar a satisfação.

Também seria interessante incorporar mais informação disponível, tal como produtos desejados e produtos já encomendados pelos utilizadores, com o intuito de tentar aumentar a percentagem de acerto nos itens relevantes. Esta informação seria bastante relevante, embora de âmbito mais reduzido, isto porque segundo os responsáveis, o sítio é preferencialmente utilizado para consultar produtos, preços e promoções, para posteriormente se deslocarem à loja física com o propósito de realizar a

aquisição, no entanto é informação que quando presente indica claramente e declaradamente uma preferência. Outro factor que afasta os utilizadores da utilização destas duas funcionalidades é a necessidade de registo.

Poderiam ainda ser implementados no sítio `WEB` outros meios de recolha de informação relativa à preferência dos utilizadores, como por exemplo: se o utilizador adicionou a página aos favoritos, se imprimiu, se sublinhou, se guardou a página no PC [15].

Outra possibilidade bastante interessante seria adicionar algumas características dos `Sistemas de Recomendação` baseados no conteúdo, de forma a obter um sistema híbrido, colmatando algumas das debilidades dos sistemas baseados em *Collaborative Filtering*. Poderia fazer-se isso mantendo o perfil relativo aos itens que cada utilizador já avaliou ou já clicou, resolvendo problemas de escassez de itens comuns e o problema dos novos itens (pois podem ser sugeridos com base nas características dos itens existentes no perfil dos utilizadores). Pode-se recorrer ainda a *bots* que simulam determinados perfis padrão conhecidos, proporcionando melhores recomendações aos utilizadores que têm preferências semelhantes (ver 2.3.2).

Bibliografia

[1] Cooley, R., Mobasher, B., and Srivastava, J., Data Preparation for Mining World Wide Web Browsing Patterns, Journal of Knowledge and Information Systems 1(1), p. 55-32, 1999.

[2] Baglioni, M., Ferrara, U., Romei, A., Ruggieri, S., and Turini, F., Preprocessing and Mining Web Log Data for Web Personalization, 8th Italian Conf. on Artificial Intelligence, Vol. 2829 of LNCS, p. 237-249, Itália, 2003. Link para o artigo

[3] Cooley, R., Tan, P. N., and Srivastava, J., Discovery of Interesting Usage Patterns from Web Data, Revised Papers from the International Workshop on Web Usage Analysis and User Profiling, p.163-182, 1999. Link para o artigo

[4] Nasraoui, O., World Wide Web Personalization, University of Louisville, USA, 2004. Link para o artigo

[5] Wei, C. P., Shaw, M. J., and Easley R. F., A SURVEY OF RECOMMENDATION SYSTEMS IN ELECTRONIC COMMERCE, University of Illinois, USA, 2003. Link para o artigo

[6] Breese, J. S., Heckerman, D., and Kadie, C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), p. 43-52, Morgan Kaufman, San Francisco, USA, 1998. Link para o artigo

[7] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon L., and Riedl J., GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM, v.40, n.3, p.77-87, 1997. Link para o artigo

[8] Konstan J., Miller B., Riedl J., Experiences with GroupLens: Making Usenet Useful Again, Proceedings of the 1997 Usenix Winter Technical Conference, 1997. Link para o artigo

[9] Domingues, M., An Independent Platform for Monitoring, Analysis and Adaptation of Web Sites, Proceedings of the 2008 ACM conference on Recommender systems, p. 299-302, Lausanne, Switzerland, 2008.

[10] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., Item-based Collaborative Filtering Recommendation Algorithms, Proceedings of the 10th international conference on World Wide Web, p. 285 - 295, Hong Kong, Hong Kong, 2001. Link para o artigo

[11] Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, p. 207 - 216, Washington, D.C., USA, 1993. Link para o artigo

[12] Adomavicius, G., Tuzhilin, A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, v.17 n.6, p. 734 - 749, 2005. Link para o artigo

[13] HERLOCKER, J. L., Understanding and Improving Automated Collaborative Filtering Systems. Tese (Doutoramento em Ciência da Computação), Universidade de Minnesota, Minnesota, USA, 2000. Link para o artigo

[14] Shardanand, U., Maes, P., Social Information Filtering: Algorithms for Automating "Word of Mouth", Proceedings of the SIGCHI conference on Human factors in computing systems, p. 210 - 217, Denver, Colorado, USA, 1995. Link para o artigo

[15] Douglas, W. O. and Kim, J., Implicit Feedback for Recommender System, Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering, p. 31 - 36, 1998. Link para o artigo

[16] Schafer, J. B., Konstan, J. A., Riedl J., E-Commerce Recommendation Applications, Data Mining and Knowledge Discovery, v. 5 n. 1-2, p. 115 - 153, 2001. Link para o artigo

[17] Gama, J., Árvores de decisão, *Power Point* da disciplina de Extracção de conhecimento de dados, 1º semestre, Universidade do Porto, Faculdade de Economia, Portugal, 2002.

[18] Alspector, J., Kolcz, A. and N. Karunanithi, Feature-based and Clique-based user models for movie selection: A comparative Study, User Modeling and User-Adapted Interaction, v.7 n.4, p. 279 - 304, 1997. Link para o artigo

[19] Raymond ,J. M. , Roy L., Content-Based Book Recommending Using Learning for Text Categorization, Proceedings of the fifth ACM conference on Digital libraries, p. 195 - 204, San Antonio, Texas, USA, 2000. Link para o artigo

[20] Soboro, I. M. and Nicholas, C. K., Combining Content and Collaboration in Text Filtering, In IJCAI'99 Workshop: Machine Learning for Information Filtering, 1999. Link para o artigo

[21] Basu, C., Hirsh, H. and Cohen, W., Recommendation as Classification: Using Social and Content-Based Information in Recommendation,Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, p. 714 - 720, Madison, Wisconsin, USA, 1998. Link para o artigo

[22] A web server log file sample explained, http://www.jafsoft.com/searchengines/log_sample.html, último acesso: 23 de Novembro de 2008.

[23] http://en.wikipedia.org/wiki/Precision_and_recall, último acesso: 27 de Dezembro de 2008.

[24] www.amazon.com, último acesso: 11 de Janeiro 2009.

[25] http://movielens.umn.edu/html/tour/index.html, último acesso: 11 de Janeiro 2009.

[26] http://ebay.about.com/od/gettingstarted/a/gs_whatisebay.htm, último acesso: 11 de Janeiro 2009.

[27] http://www.drugstore.com, último acesso: 11 de Janeiro 2009.

[28] Recommender Systems and Collaborative Filtering, http://www.deitel.com/ResourceCenters/Web20/RecommenderSystems/RecommenderSystemsandCollaborativeFiltering/tabid/1318/Default.aspx, último acesso: 13 de Janeiro 2009.

[29] Platform for Privacy Preferences (P3P) Project, http://www.w3.org/P3P/, último acesso: 15 de Janeiro 2009.

[30] C# (C Sharp), http://en.wikipedia.org/wiki/C_Sharp_(programming_language), último acesso: 2 de Fevereiro 2009.

[31] SQL Server, http://en.wikipedia.org/wiki/Microsoft_SQL_Server#SQL_Server_2005, último acesso: 2 de Fevereiro 2009.

[32] Microsoft Visual Studio, http://en.wikipedia.org/wiki/Microsoft_Visual_Studio, último acesso: 2 de Fevereiro 2009.

[33] `Internet` world stats, http://www.internetworldstats.com/stats.htm, último acesso: 14 de Julho 2009.

[34] ecommerce-nielsen, http://www.buzzes.eu/blogit/2008/01/ecommerce-nielsen.html, último acesso: 16 de Julho 2009.

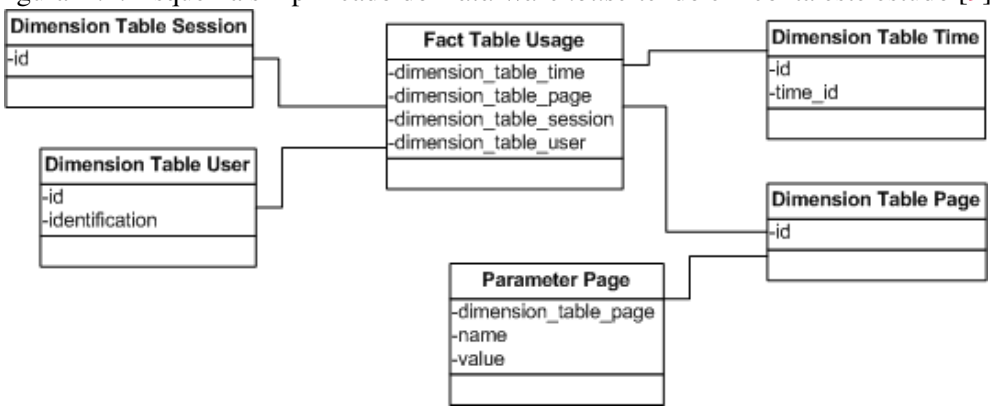
Apêndice A

Detalhe da preparação dos dados

A.1 *Data Warehouse*

Nesta ponto pretende-se realizar uma breve descrição das tabelas do *Data Warehouse* utilizadas para obter os conjuntos de dados utilizados neste estudo. Os dados existentes no *Data Warehouse* foram obtidos a partir do tratamento dos *Web Access Logs*, onde se procedeu ao pré-processamento e recolha de toda a informação associada. Toda esta informação é catalogada e colocada nos diferentes factos e dimensões do *Data Warehouse*. A figura A.1 apresenta um esquema simplificado do *Data Warehouse*, onde são especificados os atributos mais importantes para a obtenção do conjunto de dados para este estudo.

Figura A.1: Esquema simplificado do *Data Warehouse* tendo em conta este estudo [9].



As dimensões utilizadas neste estudo foram *User* (Utilizador), *Session* (Sessão), *Time* (Tempo) e *Page* (Página), com a respectiva sub-dimensão *Parameter Page* (Parâmetros da página). A tabela de factos utilizada é *Usage* (Utilização).

A dimensão *Utilizador* permite identificar os diferentes tipos de visitantes do sítio `WEB`. O campo 'identification' guarda informação relativa ao IP (Internet Protocol) ou nome do servidor (Internet Service Provider). A relação com o facto Utilização é realizada através do atributo 'id'.

A dimensão *Sessão* é constituída pelos acessos que determinado utilizador realizou ao sítio `WEB`, isto quer dizer que engloba todas páginas visitadas durante determinado período de tempo. Por este motivo um utilizador pode ter mais que uma sessão. Para relacionar as duas dimensões é necessário utilizar a tabela dos factos, através do atributo 'id'.

A dimensão *Tempo* guarda a informação relativa ao tempo em que determinada página foi acedida. Para relacionar esta dimensão com as restantes é necessário faze-lo utilizando a tabela de factos *Utilização*.

A dimensão *Página* guarda o URI (Uniform Resource Identifier) único que identifica cada página. Também para esta dimensão a relação com as restantes terá que ser por intermédio da tabela *Utilização*.

A sub-dimensão "Parâmetros da página" guarda o nome e os valores dos parâmetros associados a cada URI.

Por fim, temos a tabela de factos *Utilização*, que guarda todos os acessos de determinado utilizador ao sítio. Cada registo desta tabela apresenta relação com as dimensões apresentadas anteriormente.

A.2 Identificação de *bots*

Muitos dos *bots* são facilmente identificados através do nome do servidor de acesso ás diferentes páginas dos produtos. As palavras mais utilizadas para identificação são: 'crawl' e 'bot'. Através de uma pergunta que considere estas palavras em qualquer posição dentro do nome do servidor, foram identificados 868 *bots* que correspondem a 2.6% dos 33543 IP's e nomes de servidores. A pergunta utilizada para os identificar é a seguinte:

```
select count(a.ID)
from dw4suprides.dimension_table_user a
where a.ID in (select b.ID from dw4suprides.dimension_table_user b
where b.identification like '%crawl%'
union select c.ID from dw4suprides.dimension_table_user c
where c.identification like '%bot%');
```

Com o intuito de obter as sessões relacionadas com os utilizadores que representam *bots* é necessário utilizar a dimensão utilizador ('user'), dimensão das sessões 'session' e a tabela de factos 'Usage'. Como resultado obtiveram-se 2979 sessões.

Outra informação importante consiste na identificação da quantidade de produtos acedidos nestas sessões correspondentes a *bots*. Para tal relacionaram-se, para além das dimensões e tabela referida no paragrafo anterior, a dimensão 'page' e sub-dimensão 'parameter page'. A primeira é constituída por todas as paginas acedidas no *Web Access Logs*; a sub-dimensão representa os parâmetros utilizados no acesso ás páginas. Destes parâmetros só nos interessa o 'pid', pois é utilizado para aceder aos produtos. Posto isto, obtiveram-se 2129 produtos acedidos em sessões pertencentes a *bots*. O facto de obtermos menos acessos a produtos que sessões está relacionado com a questão de existirem sessões onde não foi realizado nenhum acesso a produtos e também porque existem outros acessos que não a produtos.

Neste processo de remoção de utilizadores e sessões é necessário considerar o administrador do sítio, pois tem um comportamento diferente dos utilizadores que pretendem realizar compras, como tal é preciso identificar o seu utilizador e respectivas sessões. A partir da análise dos *Web Access Logs* verificou-se que os acessos do administrador podem ser identificados através do parâmetro do utilizador, o que consequentemente permite obter o IP de acesso, o que possibilita por sua vez obter através do *Data Warehouse* todas as sessões associadas. Posto isto, constatou-se a existência de 1889 sessões.

Para além das operações realizadas anteriormente procedeu-se ainda a uma última validação que visa identificar se permaneceram alguns *bots* nos dados. Para tal construiu-se uma pergunta em SQL para obter o número de produtos vistos em cada sessão, de modo a identificar sessões com elevado número de produtos, pois estas constituem muito provavelmente sessões de *bots*. Este tipo de sessões

podem conduzir, quando consideradas, a alteração dos resultados. Encontrou-se uma sessão com 289 produtos visitados, o que é um número bastante exagerado, como tal não será considerada.

A.3 *Query* SQL de obtenção dos dados

Para execução do estudo apresentado nesta dissertação, utilizou-se uma abordagem que recorre a dados obtidos de forma implícita, para previsão das preferências dos utilizadores. Tipicamente os dados considerados nestas situações são do tipo binário: 1 se determinado item foi visitado e 0 senão. Neste trabalho não foi seguida esta abordagem mas sim uma abordagem em que se considerou que o facto de um utilizador visitar mais que uma vez um item na mesma sessão é demonstrativo da sua preferência.

Para obtenção dos dados utilizados como conjunto de treino e teste, tendo em conta a abordagem referida no paragrafo anterior, teve que se realizar uma pergunta ao *Data Warehouse* apresentado no ponto 3.4. A pergunta utilizada foi:

```
select d.id,c.value, count(c.value)
from dw4suprides.dimension_table_page a,
dw4suprides.fact_table_usage b,
dw4suprides.parameter_page c,
dw4suprides.dimension_table_session d,
dw4suprides.dimension_table_user e
where d.id = b.dimension_table_session
and a.id = b.dimension_table_page
and c.dimension_table_page = a.id
and c.name = 'pid'
and e.id = b.dimension_table_user
and d.id not in (21595)
and e.id not in (select b.id
                 from dw4suprides.dimension_table_user b
                 where b.identification like '%crawl%'
                 union select c.id from dw4suprides.dimension_table_user c
                 where c.identification like '%bot%'
                 union select d.id from dw4suprides.dimension_table_user d
                 where d.identification = '89.155.58.32')
group by d.id,e.identification,c.value order by d.id desc
```

Como se pode constatar, foram integradas perguntas auxiliares com o intuito de remover os *bot's* que foram detectados no ponto anterior. Verifica-se também a utilização das dimensões, sub-dimensão e atributos. É de salientar ainda o facto da utilização da condição $c.name =' pid'$, pois é através desta que se especifica qual o parâmetro que identifica o item. Ou seja, quando uma página referente a determinado item é acedida o parâmetro $'pid'$ identifica qual o item a considerar na página. Em baixo é apresentado um exemplo de acesso a um item.

```
http://www.suprides21.pt/produto.php?pid=6189
```

Para além dos itens, também se retirou informação relativa às categorias e subcategorias acedidas em cada sessão. Para estes casos bastou substituir o parâmetro utilizado, tendo-se alterado o parâmetro $'pid'$ para $'cid'$ no caso das categorias e para $'sid'$ no caso das subcategorias.

Relativamente à obtenção do tempo, procedeu-se à integração da dimensão *Tempo* de forma a obter para cada acesso a hora, minuto e segundo em que ocorreu. Com esta informação pode-se calcular a diferença entre dois acessos e assim obter o tempo de visualização de determinado item. Se na mesma

sessão ocorrer mais que um acesso a um item, procede-se então ao calculo da média do tempo de visualização.

A pergunta apresentada a seguir, permite obter os acessos e respectivos parâmetros que ocorreram durante determinada sessão.

```
select d.id, c.value,time_id,c.name
from dw4suprides.dimension_table_page a,
dw4suprides.fact_table_usage b,
dw4suprides.parameter_page c,
dw4suprides.dimension_table_session d,
dw4suprides.dimension_table_user e,
dw4suprides.dimension_table_time f
where d.id = b.dimension_table_session
and a.id = b.dimension_table_page
and c.dimension_table_page = a.id
and e.id = b.dimension_table_user
and f.id = b.dimension_table_time
and d.id in (
              select d.id as id
              from dw4suprides.dimension_table_page a,
              dw4suprides.fact_table_usage b,
              dw4suprides.parameter_page c,
              dw4suprides.dimension_table_session d,
              dw4suprides.dimension_table_user e
              where d.id = b.dimension_table_session
              and a.id = b.dimension_table_page
              and c.dimension_table_page = a.id
              and c.name = 'pid'
              and e.id = b.dimension_table_user
              and d.id not in (21595)
              and e.id not in (
                               select b.id
                               from dw4suprides.dimension_table_user b
                               where b.identification like '%crawl%'
                               union
                               select c.id
                               from dw4suprides.dimension_table_user c
                               where c.identification like '%bot%'
                               union
                               select d.id
                               from dw4suprides.dimension_table_user d
                               where d.identification = '89.155.58.32')
              group by d.id)
and d.id not in (21595)
and e.id not in (
                 select b.id
                 from dw4suprides.dimension_table_user b
                 where b.identification like '%crawl%'
                 union
                 select c.id
                 from dw4suprides.dimension_table_user c
                 where c.identification like '%bot%'
                 union select d.id
                 from dw4suprides.dimension_table_user d
                 where d.identification = '89.155.58.32')
group by d.id,e.identification,time_id,c.name, c.value order by d.id desc;
```

Para calcular o tempo de visualização de determinado item é necessário identificar o parâmetro $'pid'$ e o acesso seguinte com tempo de acesso distinto, para calcular a diferença. O pormenor do tempo de acesso distinto está relacionado com o facto de os parâmetros com o mesmo tempo serem oriundos da mesma página.

Apêndice B

Detalhes da implementação

B.1 Base de dados

Em seguida pretende-se realizar a apresentação, descrição e explicação das tabelas constituintes da base de dados utilizada na implementação.

O diagrama apresentado na figura B.1 apresenta a organização dos dados relativos aos itens, categorias, subcategorias e respectivas sessões. A tabela "session_prod_cf_cat" contem informação relativa aos acessos em cada sessão às categorias. Esta tabela é constituída por um identificador único para cada registo "s_cat_id". O campo "sessionId" representa o identificador da sessão, a que correspondem as categorias acedidas. O campo "n_value" corresponde ao valor de vezes que determinada categoria foi acedida nesta sessão. O campo "'value"' representa o identificador único da categoria.

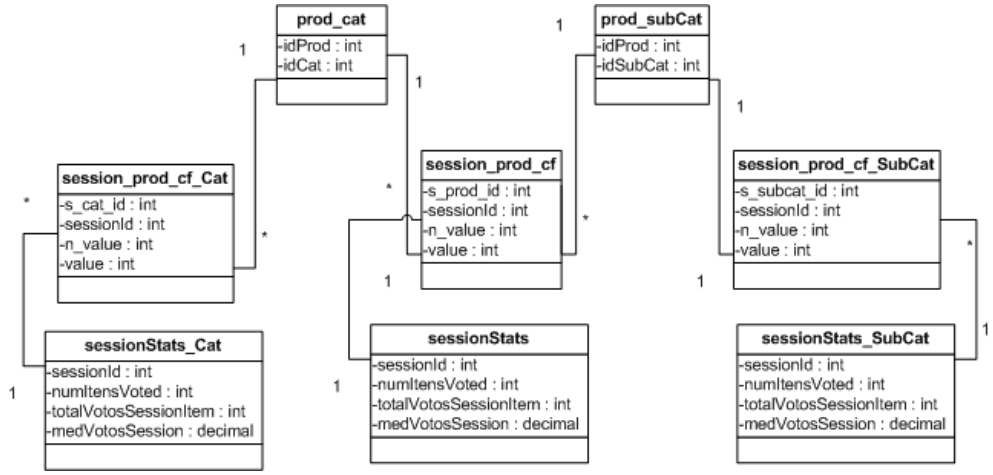


Figura B.1: Diagrama Base de Dados Itens, Categorias e Subcategorias

A tabela "sessionStats_Cat" é utilizada para conter informação relativa ao calculo da média dos acessos às categorias de determinada sessão. O campo "sessionId" corresponde ao identificador da sessão e permite relacionar esta tabela com a apresentada anteriormente. A cada sessão corresponde um registo nesta tabela. O campo "numItensVoted" identifica o número de itens distintos acedidos na sessão. O campo "totalVotosSessionItem" destina-se ao somatório de todos os acessos (votos) da sessão, incluindo os repetidos. Por último, temos o campo "medVotosSession" que corresponde ao valor da média dos votos na respectiva sessão.

Esta estrutura foi transposta para albergar os dados relativos aos itens e subcategorias.

A tabela "prod_cat" permite estabelecer a relação entre os itens e a respectiva categoria a que pertencem. A tabela "prod_subCat" é relativa às subcategorias e tem a mesma função que a tabela anterior.

O diagrama apresentado na figura B.2 é constituído pelas tabelas utilizadas para guardar informação relativa ao tempo.

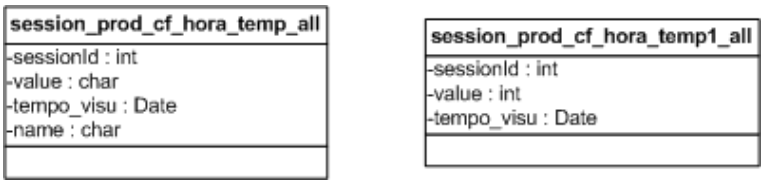


Figura B.2: Diagrama Base de Dados Tempo

A tabela "session_prod_cf_hora_temp_all" possilita guardar os dados retirados do *Datawarehouse*. Nesta são guardados os dados com todos os acessos dentro de cada sessão, com a respectiva referência temporal de quando foi feito o acesso. O campo "sessionId" constitui o identificador da sessão. O campo "value" tem como intuito reter o valor dos parâmetros do URL, ou seja, no URL "?pid=4444" este campo corresponde ao valor '4444' e o campo "name" corresponde ao nome do parâmetro, que neste caso é 'pid'. O campo "tempo_visu" contem o tempo em que ocorreu o acesso.

A tabela "session_prod_cf_hora_temp1_all" tem como intuito conter informação relativa ao tempo depois de obtido o intervalo de visualização. A partir dos dados contidos na tabela referida anteriormente, procedeu-se há obtenção do tempo de visualização de cada item, para cada sessão. Para tal criou-se um método que verifica quando o campo "name" é igual a "pid" (o que significa que se está na presença de um item) e subtrai a hora de acesso ao item pela hora de acesso ao conteúdo imediatamente posterior ao item. Posto isto, é então criado um novo registo na tabela "session_prod_cf_hora_temp1_all" contendo o identificador da sessão "sessionId", o identificador do item "value" e o resultado da subtracção mencionada anteriormente no campo "tempo_visu". Para que este método seja eficaz é necessário ordenar os dados da tabela "session_prod_cf_hora_temp1_all"por "sessionId" e "tempo_visu".

O diagrama apresentado na figura B.3 é constituído pelas tabelas utilizadas para guardar informação relativa às recomendações/previsões e realizar a avaliação.

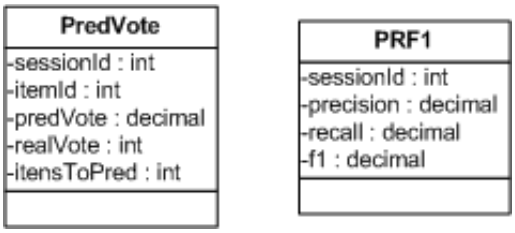


Figura B.3: Diagrama das tabelas utilizadas para avaliação

A tabela "PredVote" permite guardar os valores previstos para os diferentes itens recomendados. No campo "sessionId" é guardado o identificador da sessão, no campo "itemId" o identificador do item e no campo "predVote" o valor previsto calculado. O campo "realVote" contém o valor real do item, isto é, o valor associado a este item existente na sessão de teste para a qual se estão a gerar recomendações. O campo "itensToPred" tem como função guardar o valor de itens relevantes existentes na sessão em causa, por exemplo, se em determinada sessão de teste foram visualizados 5 itens e se for utilizada a

técnica "Dados 2 itens", então o valor deste campo é 3, pois 2 são utilizados como conhecidos. Então este campo será posteriormente utilizado para realizar a avaliação.

A tabela "PRF1" contem informação relativa à aplicação das metodologias de avaliação, obtidas a partir dos dados existentes na tabela apresentada no paragrafo anterior. Para cada sessão é gerado apenas um registo nesta tabela. O campo "sessionId" contém o identificador da sessão, o campo "precision" o resultado do calculo da medida de avaliação *Precision*, o campo "recall" contém o resultado do calculo da medida de avaliação *Recall* e por último o campo "f1" contém o resultado do calculo da medida de avaliação *F1*.

Os resultados apresentados no capítulo 4 são obtidos a partir desta tabela, através da utilização da pergunta apresentada em baixo, que calcula a média das avaliações efectuadas para cada sessão.

```
SELECT AVG(precision) * 100, AVG(recall) * 100, AVG(f1) * 100 FROM PRF1
```

B.2 Procedimentos

Neste ponto pretende-se proceder à apresentação, descrição e explicação dos procedimentos implementados.

O procedimento "calculaSugestoes" genericamente possibilita o calculo das recomendações utilizando a correlação, *Default Vote* e *Inverse User Frequency*, com ou sem tempo. É claro que os algoritmos e métodos são implementados noutros procedimentos chamados a partir deste. Os resultados obtidos por este método são inseridos na tabela "PredVote". O procedimento tem como parâmetros de entrada o número de itens a utilizar para realizar a recomendação, o tamanho do conjunto de treino, o nome da tabela onde devem ser inseridos os resultados, o nome da tabela que contém os dados relativos às sessões, itens e votos, nome da tabela da média dos votos, um booleano que especifica se o parâmetro experimental é "All but 1" e por fim o número máximo de vizinhos que se quer utilizar para calcular as recomendações.

Em seguida são enumeradas de forma sequencial as operações realizadas por este procedimento:

1. Obtém sessões de teste que obedecem ao critério resultante do somatório do parâmetro de entrada "número de itens" com 2, de forma a garantir que há pelo menos 2 itens não conhecidos, que são considerados itens relevantes aquando da avaliação.

   Para cada sessão de teste:

   (a) Obtém as sessões do conjunto de treino que visualizaram os itens conhecidos da sessão de teste;

   (b) Obtém frequência dos itens conhecidos;

   (c) Chama o procedimento "calculaMelhoresVizinhosActiveSession" com a finalidade de obter os melhores vizinhos.

   (d) Obtém itens associados às sessões dos melhores vizinhos, excluindo os itens conhecidos.

   (e) Calcula $p_{a,j}$

   No caso do tempo e dependendo da abordagem seguida na escolha dos melhores itens (ponto 4.6):

   i. Se $v_{i,j}$ igual a v_i, então o termo $(v_{i,j} - v_i)$ é substituído pelo valor/utilidade do tempo. Para tal chama o procedimento "calcTimeUtility";

   ii. Independente de $v_{i,j}$ igual a v_i, substituí $(v_{i,j} - v_i)$ pelo valor/utilidade do tempo. Para tal chama o procedimento "calcTimeUtility";

iii. Independente de $v_{i,j}$ igual a v_i, soma-se o valor/utilidade do tempo com $(v_{i,j} - v_i)$. Para tal chama o procedimento "calcTimeUtility".

O procedimento "calculaMelhoresVizinhosActiveSession", tem como objectivo a obtenção dos melhores vizinhos e tem como parâmetros de entrada: as sessões com itens iguais aos itens conhecidos do utilizador, o identificador da sessão de teste para a qual se estão a calcular os melhores vizinhos, o número máximo de vizinhos a considerar, frequência dos itens conhecidos e tamanho do conjunto de treino. Este procedimento encontra-se organizado da seguinte forma:

1. Para cada sessão com itens iguais aos itens conhecidos foram implementados 3 métodos alternativos:
   (a) Calcula o peso utilizando a correlação com/sem *Inverse User Frequency*.
   (b) Calcula o peso utilizando a correlação com/sem *Inverse User Frequency* "calculaPesoSession" caso se verifique $v_{a,j} \neq \overline{v_{a,j}}$ e para os restantes casos utilizar o *Default Vote* "calculaPesoSessionDefaultVote".
   (c) Calcula o peso utilizando a correlação com/sem *Inverse User Frequency* caso se verifique $v_{a,j} \neq \overline{v_{a,j}}$ e/ou $v_{i,j} \neq \overline{v_{i,j}}$ e para os restantes casos utilizar o *Default Vote*.
2. Se o número máximo de vizinhos a considerar já foi atingido, então procede-se à remoção do vizinho com peso inferior ao peso do vizinho encontrado nesta iteração.

O procedimento "calculaPesoSession" implementa a correlação com ou sem *Inverse User Frequency*. Tem como argumentos os dados relativos à sessão activa, os dados relativos à sessão a comparar, a especificação das colunas da tabela que contém os dados, o tamanho do conjunto de treino e os dados relativos à frequência dos itens, para o caso do calculo da correlação envolver a *Inverse User Frequency*.

O procedimento "calculaPesoSessionDefaultVote" implementa o algoritmo *Default Vote*. Tem como argumentos os dados relativos à sessão activa, os dados relativos à sessão a comparar e especificação das colunas da tabela que contém os dados.

Estes dois procedimentos foram implementados de forma a utilizarem informação relativa ao tempo, substituindo a informação relativa aos votos (calculaPesoSessionTime e calculaPesoSessionDefaultVoteTime).

O procedimento "calcTimeUtility" tem como intuito obter a utilidade do intervalo de tempo que o utilizador permaneceu a visualizar determinado item (cujo valor já foi pré calculado). Se o valor do intervalo de tempo for menor que 0,15 retorna 0, caso contrário utiliza o procedimento "utilityFunc" para obter o valor ou a utilidade do tempo dependendo da abordagem especificada. Neste procedimento também são implementas as funções de utilidade.

O procedimento "calculaSugestoesSubCat" genericamente possibilita o calculo das recomendações utilizando a correlação, *Default Vote* e *Inverse User Frequency*, informação relativa aos itens, Categorias ou Subcategorias, com ou sem tempo. Os algoritmos e métodos são implementados noutros procedimentos chamados a partir deste. Os resultados obtidos por este método são inseridos na tabela "PredVote". O procedimento tem como parâmetros de entrada: o número de itens a utilizar para realizar a recomendação, o tamanho do conjunto de treino, o nome da tabela onde devem ser inseridos os resultados, o nome da tabela que contém os dados relativos às sessões, itens e votos, nome da tabela da média dos votos, um booleano que especifica se o parâmetro experimental é "All but 1", o número máximo de vizinhos que se quer utilizar para calcular as recomendações e por fim um booleano utilizado para indicar se o conjunto de dados a utilizar é relativo às categorias ou subcategorias.

Em seguida são enumeradas de forma sequencial as operações realizadas por este procedimento:

1. Obtém sessões de teste que obedecem ao critério resultante do somatório do parâmetro de entrada "número de itens" com 2, de forma a garantir que há pelo menos 2 itens não conhecidos, que são considerados itens relevantes aquando da avaliação.

   Para cada sessão de teste:

   (a) Obtém as sessões do conjunto de treino que visualizaram os itens conhecidos da sessão de teste;

   (b) Obtém frequência dos itens conhecidos;

   (c) Chama o procedimento "calculaMelhoresVizinhosActiveSession" com o intuito de obter os melhores vizinhos;

   (d) Obtém sessão do conjunto de dados das categorias ou subcategorias com mesmo identificador da sessão de teste;

   Se existir sessão com o mesmo identificador:

   i. Obter identificadores de categorias ou subcategorias associadas aos itens conhecidos.

   ii. Obter sessões do conjunto de dados das categorias e subcategorias e respectivas estatísticas, tendo em conta as sessões comuns ao conjunto de treino dos itens.

   iii. Obter frequência das categorias ou subcategorias identificadas no ponto (i), no conjunto de dados respectivo. Tendo em conta as sessões comuns ao conjunto de treino dos itens.

   iv. Calcula melhores vizinhos através do procedimento "calculaMelhoresVizinhosActiveSession" e os dados obtidos no ponto (ii) e (iii).

   v. Compara vizinhos obtidos através dos itens com os vizinhos obtidos através das categorias ou subcategorias e escolhe melhores "getCompTable". Ou escolhe os melhores 5 vizinhos de cada abordagem "getBest5EachTable".

   (e) Obtém itens associados às sessões dos melhores vizinhos;

   (f) Calcula $p_{a,j}$

   No caso do tempo e dependendo da abordagem seguida na escolha dos melhores itens (ponto 4.6):

   i. Se $v_{i,j}$ igual a v_i, então o termo $(v_{i,j} - v_i)$ é substituído pelo valor/utilidade do tempo. Para tal chama o procedimento "calcTimeUtility";

   ii. Independente de $v_{i,j}$ igual a v_i, substituí $(v_{i,j} - v_i)$ pelo valor/utilidade do tempo. Para tal chama o procedimento "calcTimeUtility";

   iii. Independente de $v_{i,j}$ igual a v_i, soma-se o valor/utilidade do tempo com $(v_{i,j} - v_i)$. Para tal chama o procedimento "calcTimeUtility".

Para o calculo das métricas de avaliação foi implementado um procedimento ("calculaPrecisionAndRecall"), com duas especificações: uma que utiliza todos os itens sugeridos para proceder à avaliação e outra que considera os n melhores itens para realizar a avaliação. Têm como parâmetros de entrada o nome da tabela que contém os dados resultantes do processo de recomendação ("PredVote"), a tabela onde devem ser inseridos os valores das métricas ("PRF1") e no segundo caso o número de itens a recomendar.

B.3 Tecnologia

A implementação deste `Sistema de Recomendação` foi realizada utilizando a linguagem de programação C# (C Sharp), *SQL Server 2005* e *Visual Studio 2005*. A utilização da plataforma *Microsoft .Net* surgiu com o intuito de testar o seu comportamento e facilidade, para levar a cabo o desenvolvimento de `Sistemas de Recomendação`. Sistemas anteriores foram tipicamente implementados utilizando a plataforma *Java Standard Edition Development Kit* (JDK). O sistema de gestão de base de dados utilizado foi escolhido apenas por uma questão prática, porque já vem integrado no *Visual Studio 2005*, facilmente se consegue utilizar outro sistema de gestão de base de dados.

Em seguida será apresentada uma descrição mais detalhada das tecnologias mencionadas anteriormente.

A linguagem de programação C# (C Sharp) é uma linguagem orientada a objectos e faz parte da plataforma *.Net*. A sua origem baseia-se nas linguagens C++ e Java. Anders Hejlsberg é responsável por liderar a equipa de desenvolvimento do C#, inicialmente foi designada por "Cool", abreviatura de "C like Object Oriented Language". No entanto, em Julho de 2000, após a *Microsoft* colocar este projecto como publico, foi-lhe alterada a designação para C# por razões de *trademark* [30].

Os objectivos de *design* do C# foram [30]:

- Ser simples, moderna, generalista e orientada a objectos;
- Apresentar boas capacidades de verificação, como limites de vectores, detecção de tentativas de utilização de variáveis não inicializadas, portabilidade do código e *garbage collection* automático. Pretende-se assim garantir que o *software* é mais robusto, duradouro e de fácil implementação.
- Possibilitar o desenvolvimento de componentes de *software* para ambientes distribuídos.
- Ser de fácil aprendizagem, principalmente para programadores familiarizados com o C, C++ e JAVA.
- Suportar Internacionalização.
- Possibilitar a utilização quer em *software* unicamente desenvolvido nesta linguagem quer em sistemas que utilizam apenas alguns componentes desenvolvidos nesta linguagem.
- Tentar minimizar a utilização dos recursos de *Hardware*, quer ao nível da memória, quer ao nível do processador, embora não tenha por intuito competir com a performance e tamanho proporcionado por linguagens como C e Assembler.

No que diz respeito ao sistema de gestão de base de dados foi utilizado o *SQL Server 2005*. O código base para este sistema é originário da *Sybase SQL Server* e constituiu assim a entrada da *Microsoft* no mercado das base de dados. Originalmente a *Microsoft*, *Sybase* e *Ashton-Tate* trabalharam em conjunto para criar a primeira versão do *SQL Server*, por volta de 1989. Aquando do lançamento do *Windows NT* a *Microsoft* e a *Sybase* separaram-se, embora a *Microsoft* mantive-se o exclusivo dos direitos para todas as versões do *SQL Server* para o sistema operativo *Windows*. Actualmente apresenta grandes avanços ao nível de performance, ferramentas de IDE (Integrated Development Environment) e alguns pacotes complementares responsáveis pela extensão das suas funcionalidades.

Como por exemplo: ferramentas de ETL (Extract, Transform, and Load) para *Data Warehouse*, Servidor de relatórios, OLAP (On-line Analytical Processing), `data mining` e algumas tecnologias de mensagens (Service Broker e Notification Services) [31].

Como ferramenta de trabalho foi utilizado o *Visual Studio 2005* que disponibiliza um conjunto de ferramentas para o desenvolvimento de *software*, desde formulários para o *Windows*, sítios `WEB`, aplicações `WEB` e serviços `WEB` para as plataformas *Windows*, *Windows Mobile*, *Windows CE*, *.NET Framework*, *.NET Compact Framework* e *Microsoft Silverlight* [32].

O *Visual Studio 2005* inclui um editor de código que suporta *IntelliSense* (Auto completa o código) e *code refactoring*. Possui um *debugger* que apresenta uma dupla funcionalidade, pois aplica-se quer ao nível do código, quer ao nível da máquina. Inclui também um *designer* de formulários e permite complementar e expandir estas funcionalidades através da utilização de *plug-ins* [32].