Universidade de Aveiro Departamento de Matemática

2007

Rita Catarina Nunes Gaspar

Pesquisa de Informação em Catálogos Científicos

Universidade de Aveiro Departamento de Matemática
2007

Rita Catarina Nunes Gaspar

Pesquisa de Informação em Catálogos Científicos

dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática, realizada sob a orientação científica do Dr. Eugénio Alexandre Miguel Rocha, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

o júri

presidente

Prof. Dr.ª. Ana Maria Reis d´Ázevedo Breda
professora associada com Agregação da Universidade de Aveiro

Prof. Dr.ª. Ana Alice Rodrigues Pereira Batista
professora auxiliar da Escola de Engenharia da Universidade do Minho

Prof. Dr. Eugénio Alexandre Miguel Rocha
professor auxiliar da Universidade de Aveiro

agradecimentos

Muito do que sou hoje e do que consegui alcançar ao longo da minha vida e percurso académico devo a um conjunto de pessoas a quem tenho de agradecer por isso.
Foram essas as pessoas que de uma ou de outra forma, me apoiaram, motivaram, desafiaram a evoluir e a crescer intelectualmente.
Gostaria assim, em primeiro lugar, agradecer aos meus pais, meus mestres, Fernando Gaspar e Irene Gaspar, a quem devo tudo o que sou e o que consegui alcançar na minha vida. Agradecer pelo vosso conforto, apoio, carinho e por tudo o que representam para mim, pelos valores que me transmitiram e estabilidade familiar que me proporcionam.
Gostaria de agradecer ao Professor Eugénio Rocha. A sua orientação e dedicação a este trabalho manteve-me sempre motivada a continuar, mesmo quando os obstáculos eram muitos. O seu contributo e orientação foram um estímulo para evolui ao longo deste período de trabalho e investigação.
Ao Sérgio e Xana, meus colegas de trabalho no PmatE e sobretudo amigos, quero agradecer pelos dias em que me viram desanimada, me ouviram e deram apoio para continuar.
Gostaria de agradecer a alguns dos meus amigos que conheci na Universidade de Aveiro, e que de certa forma influenciaram o meu percurso de estudante e com quem passei grande parte da minha vida: à Juliana, à Filipa, à Susana, à Sónia, à Inês Sousa, ao Alexandre, ao Zé, ao André, ao Alexandre Mota, ao João Crespo, ao Pedro, ao Jorge, ao Bruno, ao Mário Wilson, à Marta, à Rita, à Mónica, à Ana Alegre e à Ana Vicente.
À minha família mais próxima: Dina Dias, Hugo Dias, Isabel Dias, Reinaldo Dias, António Dias (Tó Braga) e família, Rosa Gaspar e em especial à minha avó Maria José pelo seu carinho e afecto.
A ti João, pelo teu amor, amizade, palavras de coragem e pelo enorme apoio que sempre me deste. Pelo conforto e por todos os momentos que partilhastes comigo todos estes meses.
Ao meu avô, Joaquim Dias e à sua memória dedico esta dissertação. Esteja onde estiver partilhará sempre das minhas conquistas e um pouco do meu pensamento. O meu agradecimento e sorriso final vão para ele.

palavras-chave

Pesquisa de informação, catálogos científicos, estruturas de apontadores, PageRank, Hits, Salsa, grafos, cadeias de Markov, distância de Mahalanobis.

resumo

A actividade científica depende fortemente da pesquisa e consulta de literatura científica. Na actualidade, muito desse material bibliográfico e de referência encontra-se disponível na Internet. No entanto, com o rápido crescimento da rede telemática tem surgido uma dificuldade acrescida para investigadores identificarem e utilizarem toda a informação relevante ao seu trabalho. Em particular, na área da matemática o problema também começa a irromper. De facto, a literatura matemática tem uma longa tradição de organização, devido à sua natureza acumulativa e, ao contrário de outras Ciências, a importância que resultados antigos têm na investigação presente. Actualmente, a produção científica é elaborada e transmitida em forma digital, o que permite que esta possa ser disponibilizada e acedida na rede. No entanto, neste universo complexo de informação torna-se difícil classificar a relevância, para determinado utilizador/investigador, do conteúdo de certos documentos científicos. Em geral, os documentos científicos de mais fácil acesso tendem a ser mais citados. Este facto, impele os próprios investigadores a desejar que os catálogos científicos sejam mais eficientes e passíveis de fácil acesso na Internet.

Formas de pesquisa global e integrada são cada vez mais importantes e necessárias para o desenvolvimento e progresso da Ciência, em particular, na Matemática. Assim, no âmbito deste trabalho é proposto um modelo matemático para pesquisa de documentos matemáticos em catálogos científicos. Este modelo permitirá que o sistema mostre ao utilizador de forma organizada numa topologia hierárquica, os documentos mais relevantes de acordo com a sua pesquisa. Para isso, formula-se um mecanismo que calcula a afinidade entre todos os artigos de acordo com uma distância criada a partir dos códigos MSC (Mathematical Subject Classification) que os mesmos contêm. Para além disso, o sistema poderá englobar um conjunto de novas funcionalidades: permitirá, não obstante a comum pesquisa básica de artigos, que o utilizador registado tenha acesso a notícias personalizadas automaticamente e actualizadas periodicamente. Essas notícias chegarão ao conhecimento do utilizador dando-lhe conta que um novo artigo próximo da sua área científica, ou interesse declarado, foi publicado no sistema. Um utilizador/autor poderá ter acesso a uma lista de autores que têm interesses na mesma área de investigação, de certo modo, que se encontram na sua vizinhança científica.

keywords Informal Retrieval, citation database system, Hyperlinks, PageRank, Hits, Salsa, Graphs, Markov chain, Mahalanobis Distance

abstract The scientific activity strongly depends of the search and queries of bibliographic content. Nowadays that content in highly available in the Internet. With the fast increasing of the World Wide Web becomes difficult for authors, reviewers and users identify and work with all information relevant for their investigations. As well, in mathematics this problem is growing up. The reasons for this are that mathematical literacy have a long tradition of organization, in part because her accumulative nature, in part because of necessities of mathematicians to create systems of information retrieval more efficient and possible to access by web. The mathematical articles that are easier to access usually became more cited. In these days, all mathematical articles are elaborated and transmitted in a digital format, permitting their access by users in web. Although, in this complex world of information becomes difficult categorize as relevant the content of these documents, especially when we focus on all the different needs of users.

Ways of global and integrated search are each time more important and necessary for the development and progress of scientific catalogs. Therefore, in the scope of this work we propose a mathematical model for search of mathematical documents in citation database systems. This model will allow the system to show the user, in an organized way by a hierarchical topology, relevant documents related with his search. Thus, is formulate an mechanism to reckon up the analogy between all the articles according to a distance created based in MSC codes (Mathematical Subject Classification) that the same articles have.

Besides, the system may agglomerate a set of new functionalities: consent, in spite of the common search of articles, that register users have access to RSS Feeds. This feeds will reach the user, giving him the opportunity to know that relevant articles have been published in the system. Users/authors may have access to a list of authors that have interest in the same investigation area, in some sense, that are in some scientific neighborhood.

Conteúdo

1 Introdução **1**

1.1 Motivação 1

1.2 Sobre a Internet 3

1.3 Os Catálogos Científicos 4

1.4 Plano da Dissertação 5

1.5 Contribuição 6

2 Modelos de Pesquisa Baseados no Conteúdo **9**

2.1 Introdução 9

2.2 Dados e Mecanismos de Interacção 11

2.3 Métodos de Recuperação e Visualização da Informação 13

2.3.1 Modelo vectorial 13

2.3.2 Modelo booleano 16

2.3.3 Modelo probabilístico 18

2.3.4 Redes bayesianas 20

3 Algoritmos de Pesquisa Baseados em Apontadores **23**

3.1 Introdução 23

3.2 Algoritmo Hits 29

3.2.1 Implementação do algoritmo Hits 31

3.2.2 Convergência do algoritmo 33

3.2.3 Exemplo 35

3.3 Algoritmo PageRank 36
3.3.1 Existência e convergência 40
3.3.2 Número de iteracções 41
3.3.3 Critérios de convergência 42
3.3.4 Implementação do algoritmo PageRank 42
3.3.5 Exemplo 44
3.4 Algoritmo SALSA 47
3.4.1 Convergência do algoritmo 50
3.5 Análise Crítica dos Algoritmos 52
3.6 Efeito de Apontadores Inapropriados no Algoritmo PageRank 59
3.6.1 Caso geral: diversas páginas com apontadores inapropriados . . 60
3.6.2 Caso particular: uma página com apontadores inapropriados . . 60

4 Aplicação do PageRank a um Catálogo Científico 65
4.1 Mecanismo de Pesquisa 65
4.2 Pesquisa de Documentos por Aplicação do PageRank 67
4.3 Pesquisa de Autores por Aplicação do PageRank 67

5 Um Novo Modelo de Pesquisa para Catálogos Científicos 73
5.1 Pesquisa Personalizada 73
5.2 Modelo Matemático 75
5.2.1 Catálogos científicos de referência 75
5.2.2 Distância entre códigos MSC 79
5.2.3 Espaço vectorial livre gerado pelos códigos MSC 83
5.2.4 Documentos como elementos de $\mathbb{R}_{\mathtt{MSC}}$ 84
5.2.5 Autores como elementos de $\mathbb{R}_{\mathtt{MSC}}$ 85
5.3 Uma distância em $\mathbb{R}_{\mathtt{MSC}}$ 85

6 Novas Funcionalidas de um Catálogo Científico 89
6.1 Ordenação 89

6.2 Agrupamento de Metadados 90
6.3 Vizinhança de um Artigo 92
6.4 Vizinhança de um Autor/Utilizador 94
6.5 Classificação Automática de Autores 95
6.6 Notícias Personalizadas 96

7 Conclusão e Trabalhos Futuros **97**
7.1 Conclusão 97
7.2 Trabalhos Futuros 98

Capítulo 1

Introdução

1.1 Motivação

A Rede Telemática contém uma quantidade considerável e complexa de informação, mas a facilidade com que se pode pesquisar ou armazenar informação tornou-a numa das bases de dados mais importantes dos dias de hoje.

Os sistemas de acesso electrónico a fontes de informação complexas necessitam de suportar mecanismos de pesquisa de informação cada vez mais eficientes, independentes do tipo de informação que cada utilizador pretende armazenar. Em geral, a falta de organização, indexação dos dados e a repetição de informação na Rede Telemática potencia a que determinada informação não seja identificada como relevante, mesmo numa pesquisa a esta dirigida.

No processo de pesquisa de informação é relevante a proximidade entre as palavras definidas na consulta do utilizador e as palavras-chave definidas nos documentos, aumentando ou não as hipóteses de determinados documentos serem considerados relevantes para uma pesquisa.

Devido ao crescimento exponencial de informação disponível, são cada vez mais evidentes as limitações dos sistemas de pesquisa e recolha de informação, e.g., estes aglomeram

um conjunto muito vasto de informação e conduzem muitas das vezes à dispersão da informação em vários domínios. Veja-se, na área da investigação de carácter científico onde o número de documentos publicados tem aumentado de forma mais que linear ou mesmo quase exponencial, onde existe a preocupação permanente, por parte da comunidade científica e em particular nas ciências exactas (e.g. a Matemática), em desenvolver catálogos científicos mais eficientes, em que a pesquisa de dados forneça resultados mais relevante e que suportem o trabalho dos investigadores.

A MathSciNet e a Zentralblatt são os principais e mais conhecidos catálogos científicos na área da Matemática. Contudo, a crescente dificuldade em categorizar os documentos e publicações tem causado algumas limitações nestes sistemas, não dando ao utilizador a qualidade nem a relevância que necessitam nas pesquisas que efectuam.

Em particular na matemática, a investigação depende fortemente das publicações e do conhecimento desenvolvido num longo período de tempo, o que não acontece em outras ciências, e.g. na Química onde este período é geralmente de 5 anos.

Técnicas tradicionais de pesquisa e modelos de representação, tais como o Modelo Vectorial, Booleano ou Probabilístico, não serão suficientes para serem aplicados em sistemas deste tipo. Por outro lado, algoritmos como o PageRank, Hits e Salsa desenvolvem-se à volta de uma estrutura de apontadores previamente conhecida, mas que não suporta a informação relevante acerca das publicações feitas na área da Matemática.

Na comunidade Matemática, adoptou-se um sistema de classificação de documentos, baseado em códigos estruturados e pesquisáveis numa estrutura de árvore, atribuídos a cada publicação pelos seus autores e validados pelos revisores e editores de revistas. No entanto, a atribuição humana e a exaustão da revisão de classificações poderá produzir algumas falhas e levar a erros de classificação.

Nesta dissertação é apresentado um modelo que tenta ultrapassar as limitações referidas anteriormente, utilizando a classificação por códigos referida precedentemente.

Propõe-se um modelo de representação e pesquisa num catálogo de Matemática que permita desenvolver novas potencialidades de pesquisa de informação, utilizando me-

canismos de pesquisa por documento ou autor, que encaixam ambos no mesmo espaço vectorial livre, onde as métricas usadas para medir distâncias entre documentos são diferentes das usuais, nomeadamente utiliza-se a distância do tipo de Mahalanobis.
Por fim apresentam-se novas funcionalidades que derivam das potencialidades propostas, sendo estas de grande utilidade para os utilizadores de um catálogo científico.

1.2 Sobre a Internet

A Internet é a rede das redes. O seu desenvolvimento é coetâneo com o da ethernet; ou seja, a década de sessenta, embora tenha adquirido a denominação de Arpanet.
No ínicio, o principal objectivo da Internet era proporcionar comunicação entre as diferentes bases militares dos Estados Unidos, uma vez que apesar de a segunda guerra mundial ser apenas um fantasma a guerra fria era uma realidade.
Quando esta terminou, a Arpanet já não foi vista como exclusivamente útil para os militares, tendo sido por isso generalizado o seu uso a cientistas, universidades e todo o tipo de utilizadores da rede [50]. Actualmente mais de 30 milhões de pessoas estão ligadas a esta rede de comunicação mundial.
Hoje, a Internet é indubitavelmente o maior sistema de troca de conhecimento que o Homem alguma vez desenvolveu .
Por consequência, como o aparecimento da World Wide Web a própria rede foi-se aperfeiçoando, nomeadamente com a introdução de novos protocolos como o archie, gopher, ftp e o http.
Desta forma, o conhecimento presente na rede tornou-se agradável devido à inserção de novas funcionalidades, como por exemplo sons e imagens, bem como meios de localização da informação pesquisada por arquivos com um endereço próprio.
Em suma, a vantagem da Internet é proporcionar uma troca de informação e a comunicação mundial, através de protocolos e serviços, tais como: correio electrónico, grupos de discussão e de notícias, pesquisa e cópia de ficheiros e programas informáticos, exe-

cução remota de programas e serviços de informação [6].

1.3 Os Catálogos Científicos

A evolução científica e tecnológica tem favorecido o aumento do conhecimento, por um lado, e à sua fragmentação por outro, em prol do aparecimento de novas áreas do saber, evidenciado no vasto número de resultados científicos publicados em diversos suportes. Refira-se que não só a forma de apresentação do conhecimento mudou como também o seu suporte. Na actualidade, o conhecimento está registrado tanto em suportes tradicionais como em suportes electrónicos, no entanto, é maioritariamente produzido em suporte digital.

Este aumento do conhecimento traduz-se num aumento da quantidade de informação que hoje está disponivel e que favorece a ilusão de que estamos perante uma sociedade produtora e consumidora de informação eficiente.

A necessidade de se investir num tratamento técnico dos recursos informacionais, assim como na sua organização tornaram-se inevitáveis na pesquisa e recuperação de informação. Segundo esta necessidade, a catalogação da informação é importante como forma de representar o conhecimento. Nomeadamente, a catalogação da informação no âmbito de conhecimento cientifíco, isto porque a ciência está descrita como um veiculo de comunicação rápida entre a sociedade e neste ciclo de comunicação os catálogos científicos têm um papel importante na pesquisa de informação científica.

A catalogação científica pode ser vista como um conjunto de actividades que consistem em identificar nos documentos os seus descritivos ou metaproposições e, em seguida, extrair os descritores indicadores do seu conteúdo, para posterior recuperação (metadados). Esses descritores constituem-se na representação dos elementos indicadores do conteúdo de cada documento e na sua representação. Existem várias formas de catalogação [51], e.g. catalogação por palavra-chave ou termos, catalogação lexical, onde num documento é desmontado cada discurso onde as palavras têm sentido em função

do contexto, catalogação sintagma e muitas outras. A maneira de catalogar um documento depende, naturalmente, do tipo de documento a catalogar e da forma como se pretende posteriormente aceder a este. Assim, qualquer que seja a maneira de catalogação, esta deverá permitir ao utilizador o acesso eficiente ao documento que contém a informação que pretende e necessita.

1.4 Plano da Dissertação

De modo a enquadrar o modelo matemático e as novas funcionalidades, propostas nesta dissertação, ir-se-ão apresentar alguns dos modelos de representação e pesquisa conhecidos na literatura, baseados no conteúdo ou baseados em apontadores.
No próximo capítulo, é apresentada uma descrição pormenorizada dos algoritmos tradicionais para pesquisa de informação baseada no conteúdo. No capítulo 3, definem-se os algoritmos de pesquisa de informação baseados na estrutura de apontadores mais utilizados na rede telemática, tais como, o algoritmo Hits, PageRank e Salsa. Para cada um destes algoritmos é estudada a sua convergência e exemplificado a sua aplicação.
No quarto capítulo, realiza-se uma breve descrição teórica da aplicação do algoritmo PageRank a um catálogo científico, para ordenação de documentos mais citados e para ordenação de autores mais populares do catálogo científico.
No quinto capítulo, apresenta-se um novo modelo matemático de pesquisa de dados com potencial aplicação a um catálogo científico na área da Matemática. Definem-se, no capítulo 6, algumas funcionalidades e mecanismos que o sistema poderá adicionar, considerando a representação e métrica introduzidas no capítulo 5.
Por fim apresentam-se, no capítulo 7, as conclusões e algumas ideias de trabalho futuro.

1.5 Contribuição

O presente trabalho teve, como foco, a contribuição com um novo modelo matemático para a pesquisa de documentos em catálogos científicos, na área da Matemática.
Com o crescimento da Internet e de novas formas de disseminação da cultura e da informação científica tem sido necessário, por parte de investigadores, estudar e investigar novos mecanismos de pesquisa. De um modo geral, a informação que o utilizador pretende obter dos sistemas não corresponde ás suas necessidades. Também na área da matemática e para todos os que a ela recorrem ou a utilizam torna-se difícil obter resultados positivos após a pesquisa num catálogo científico.
Os documentos matemáticos são actualmente elaborados e transmitidos em formato digital, o que lhe permite integrar outros elementos para além dos tradicionais texto e imagens estáticas. Criar mecanismos de pesquisa mais avançados para catálogos científicos na área da matemática é o principal propósito deste trabalho, pois os computadores e a Internet alteraram radicalmente a nossa forma de comunicar e de partilhar ideias e resultados. No entanto, este conjunto de mudanças e adaptações ainda não estabilizou novos meios de desenvolver, partilhar e pesquisar as ciências Matemáticas. Contribuir para a necessária consciencialização e exploração de novas técnicas de pesquisa é uma das principais missões que levou à investigação e realização deste trabalho. Pretende-se contribuir com este trabalho para o melhoramento e criação de novos mecanismos de explorar a Matemática, pois a Matemática é particularmente dependente da sua literatura, da sua credibilidade e acessibilidade, que podem ser facilmente melhoradas considerando as crescentes facilidades de publicação e de comunicações electrónicas em catálogos científicos.
Este trabalho foi apresentado à comunidade científica na apresentação "Personalizing a Citation Database System by using Mathematics Subject Classification", na conferência CMDE'06 realizada em Aveiro, 15-18 de Agosto de 2006. O conteúdo da apresentação suscitou o interesse da American Mathematical Society, com a qual se está a desenvolver

um projecto tendo por base o modelo matemático apresentado nesta dissertação.

Capítulo 2

Modelos de Pesquisa Baseados no Conteúdo

2.1 Introdução

Ao longo dos tempos, a quantidade de documentos armazenados na rede telemática tem aumentado consideravelmente. Apesar da evolução das tecnologias da informação e da computação o processo de organização, estruturação e manuseamento dos documentos é ainda muito complexo e demoroso, citando O. Cardoso [16] «*A crescente complexidade dos objectos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados ...*».

Por exemplo, as bibliotecas bigitais [10], continuam hoje em dia, a ser instrumentos criados para lidar com a informação na rede telemática. Apresentam formas inovadoras de produzir conteúdo e serviços para controlar a dispersão cultural e científica da informação.

As bibliotecas bigitais não são mais do que um conjunto de serviços e sistemas informáticos controlado por uma entidade. Como o próprio nome indica acoplam informação no formato digital, que facilmente é acedida por múltiplos utilizadores, por meio, por

exemplo de "e-books" ou acesso remoto. No entanto, bibliotecas digitais e todos os outros sistemas actuais de visualização e recuperação de informação (SVRI) [25] não estão preparados para manusear na perfeição o excesso de informação na rede. Apenas, reunindo conceitos de computação, interfaces relacionais, noções de separação de dados e algoritmos matemáticos, a visualização gráfica de informação permite a apresentação dos documentos existentes no sistema em formatos digitais. Permite igualmente a percepção, a análise e compreensão dos dados por parte dos utilizadores [25].

De outro modo, pode-se constatar que os SVRI têm a finalidade de representar dados de um universo de dados, de tal forma que esta representação explore a percepção humana, isto é, explore a percepção da informação por parte do utilizador. Por sua vez, o utilizador deverá interpretar os resultados apresentados pelo sistema com a intenção de depreender novo Saber.

Assim, na construção de sistemas de visualização e recuperação gráfica de informação, é pertinente considerar tanto a melhor forma de explorar a informação como ponderar os melhores meios de representação da informação, limitando a quantidade de informação devolvida ao utilizador, aumentando a relevância da mesma. O tipo de informação a ser tratada e as acções que precisam de ser realizadas pelo utilizador são factores relevantes na escolha dos SVRI.

A existência de locais na rede telemática, como por exemplo o Altavista, Hotbot, Google, Yahoo e muitos outros locais, com o único objectivo de facilitar a pesquisa de documentos relevantes para o utilizador, surgem da popularidade que os SVRI têm actualmente. Popularidade esta devida maioritariamente ao elevado número e variedade de documentos na rede.

Por outro lado, tal como acontece em muitas das áreas do conhecimento, a documentação científica assume um papel cada vez mais importante, com carácter estratégico e decisivo na evolução da rede telemática. Esta documentação científica tem um carácter que a distingue claramente de outras na rede. Em particular pela sua especifícidade, os mecanismos de obtenção de informação requerem uma certa adaptação, utilizando apontadores e indicadores específicos.

Esta especialização altera o modelo matemático dos algoritmos utilizados nesses mecanismos, criando novas dificuldades e problemas a resolver.
Nas próximas secções apresentam-se alguns dos modelos existentes para pesquisa e representação de informação, mas para melhor entendimento do funcionamento destes modelos apresenta-se de seguida uma breve caracterização dos tipos de dados e de mecanismos de interacção a estes subjacentes.

2.2 Dados e Mecanismos de Interacção

Num sistema de informação, os dados são caracterizados como sendo entidades ou objectos que definem processos alvo de estudo e análise [30]. Considerando os dados entidades importantes na elaboração de modelos de visualização é necessário ter em conta, para além do seu modo de armazenamento nos modelos, a sua categorização para uma eficiente utilização.
A categorização pode ser definida de formas diversas, por exemplo, categorização pelo tipo de dados: vectorial, alfanumérica, escalar entre outros. Também se pode ter categorização de acordo com a natureza e dimensão do seu domínio de dados: domínio unidimensional, bidimensional ou n-dimensional contínuo ou discreto [21].
Estabelecer meios de comunicação entre o sistema e o utilizador não é tarefa fácil; é necessário criar mecanismos de interacção funcionais em diversos níveis. A representação estática da informação normalmente não é suficiente para garantir a percepção humana dos dados. Num nível inferior, são necessários mecanismos de interacção como por exemplo o "scroll bar". Num segundo nível, mecanismos de selecção de dados relevantes que podem originar reposição ou discriminação dos dados ao utilizador. Num nível superior são necessários mecanismos de selecção de subconjuntos de dados do conjunto de dados do sistema perante critérios de pesquisa por parte do utilizador [30]. Assim, ao processo de recuperação e selecção de um conjunto de dados do universo de dados do sistema, que responde a determinada consulta realizada por um utilizador,

dá-se o nome de processo de recuperação e visualização da informação. E este conjunto de dados recuperado é apresentado por ordem decrescente de relevância [15].
Um sistema de visualização e recuperação de informação pode ser representado como se mostra na figura 2.1.

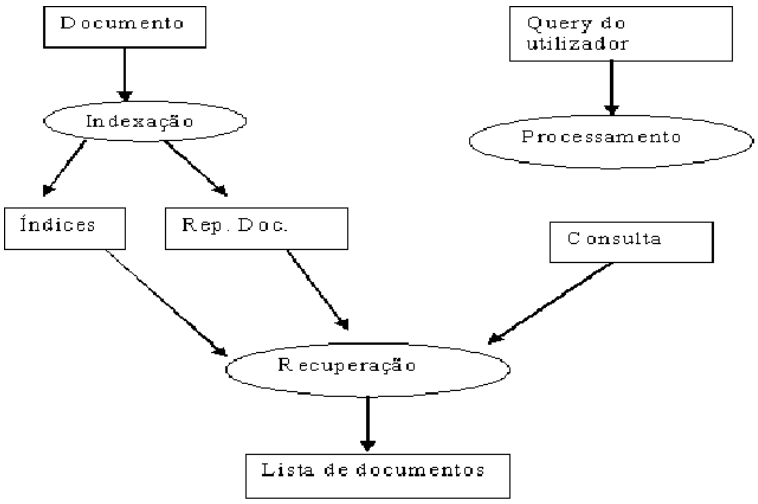


Figura 2.1: Sistema de visualização e recuperação de informação

Os modelos clássicos [15], Modelo Vectorial, Modelo Booleano e Modelo Probabilístico, foram os primeiros modelos a surgirem e a serem utilizados para consulta e recuperação de informação de documentos. Esta recuperação de informação é baseada no conteúdo dos documentos e permite a visualização da informação ordenada por ordem de relevância dos documentos. O conteúdo de cada documento é extraído através da pesquisa de termos nos documentos. Porém na rede telemática, existe uma considerável quantidade de informação que pode ser extraída da estrutura de apontadores entre documentos, que não é considerada nos modelos clássicos e será abordado no próximo capítulo deste trabalho. Em geral, combinar a extracção de informação baseada em conteúdo com a extracção de informação baseada em apontadores será a melhor estratégia a explorar para melhorar a ordenação de (meta)dados. Na próxima secção ir-se-á focar os modelos clássicos de visualização e recuperação de informação.
Além dos modelos clássicos, referidos anteriormente, outros modelos mais complexos têm sido apresentados ao longo dos anos, destacando-se entre muitos, os modelos basea-

dos em bases de conhecimento, lógica "fuzzy" e redes neurais, [15], que não serão abordados neste trabalho.

2.3 Métodos de Recuperação e Visualização da Informação

2.3.1 Modelo vectorial

O Modelo Vectorial (M.V.) foi definido por Salton em 1968, a sua aplicação é largamente difundida nas operações de categorização automática e filtração de informação em documentos. Este modelo, representa documentos e consultas ("queries") na forma de vectores de termos sendo cada documento um registro de dados que inclui uma parte textual. O M.V. permite reconhecer a relação entre as propriedades de cada documento e as propriedades de um conjunto de documentos.
Considere as seguintes definições de *ordem parcial* e *total* de um conjunto.

Definição 1 *Um conjunto X, com uma relação $\leq$ associada diz-se um conjunto com ordem parcial se satisfaz as seguinte condições: Para todo $a, b, c \in X$*

- $a \leq a$ *(reflexividade);*
- $a \leq b$ *e* $b \leq c \Rightarrow a \leq c$ *(transitividade)*
- $a \leq b$ *e* $b \leq a \Rightarrow a = b$ *(antí-simétrica)*

Este conjunto diz-se de ordem total *se é um conjunto de ordem parcial tal que* $\forall a, b \in X, a \leq b$ *ou* $b \leq a$.

Seja D uma cadeia de documentos, i.e., um conjunto de documentos juntamente com uma ordem total (ver definição 1). O j-ésimo documento de D é denotado por d_j. Um termo é uma palavra que semanticamente incorpora o conteúdo principal do documento. O conjunto dos termos é denotado por Γ.

A cada termo $t_i \in \Gamma$ pode-se associar um documento d_j e um determinado peso associado $w_{ij} \geq 0$, que quantifica a correlação entre o termo t_i e o documento d_j. Uma consulta é um conjunto de termos que, de algum modo, expressa a intenção do utilizador, e é denotado por Q. As consultas são encaradas como documentos (temporários), e como tal, têm as mesmas propriedades e estrutura destes. Deste modo, consultas e documentos são representados como vectores de um espaço N-dimensional, onde N é o número de termos do conjunto, i.e., $N \equiv |\Gamma|$. Cada elemento deste espaço é encarado como um vector de termos $t_i \in \Gamma$ ortogonal, isto é, $i \neq j \Rightarrow < t_i, t_j >= 0$, com $t_i, t_j \in \Gamma$ e $< ., . >$ denota o produto interno em $\mathbb{R}^N$. Este facto permite afirmar que estes termos ocorrem independentemente uns dos outros dentro dos documentos e consultas. Para além disso, assume-se que para qualquer $t_i \in \Gamma$, $|t_i| = 1$.

Documentos e consultas podem ser então representados por vectores da forma $v_{d_j} = (w_{1j}, w_{2j}, ..., w_{Nj})$ e $v_Q = (w_{1Q}, w_{2Q}, ..., w_{NQ})$, onde w_{ij} e w_{iQ} são pesos associados ao termo t_i no documento d_j na consulta Q respectivamente [36].

Os documentos devolvidos como resultado de uma consulta são representados similarmente, isto é, o vector resultado de uma consulta é definido através do cálculo de similaridade [17]. O cálculo da similaridade entre documentos é baseado no ângulo formado entre os vectores de representação dos documentos de acordo com a seguinte expressão

$$Sim(d_m, d_n) = \frac{\sum_{t \in \Gamma} w_{td_m} w_{td_n}}{\sqrt{\sum_{t \in \Gamma} w_{td_m}^2} \sqrt{\sum_{t \in \Gamma} w_{td_n}^2}}. \tag{2.1}$$

Geometricamente a interpretação do cálculo de similaridade entre documentos é dada como se mostra na figura 2.2, que corresponde ao cálculo do coseno do ângulo entre vectores representativos dos documentos. Em particular, pretende-se desta forma estudar (ordenar) os valores $Sim(d, q)$ para $d \in D$ e $q \in Q \subset D$.

Em concreto, os pesos associados aos termos dos documentos numa consulta são calculados fazendo o equilíbrio entre as características do documento, nomeadamente considerando a ocorrência de determinado termo no documento. Se o conjunto de documentos D possui K documentos e n_d é o número de ocorrências do termo t no

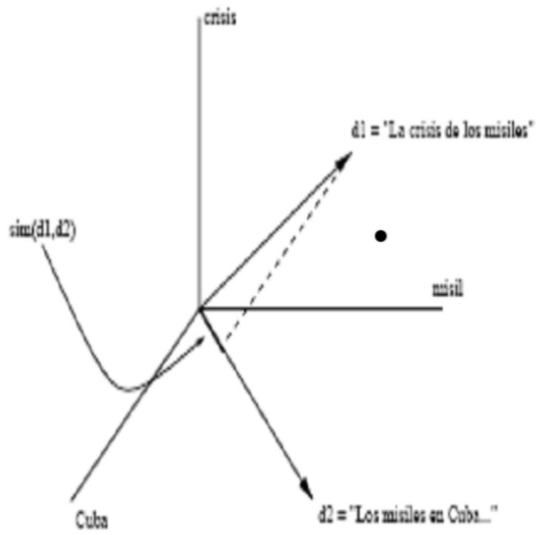


Figura 2.2: Interpretação geométrica do modelo vectorial, onde d_1 e d_2 representam documentos e $\text{sim}(d_1, d_2)$ a similarideade entre eles

documento d então o peso associado a $d \in D$ é dado pela expressão

$$w_d = freq(t, d) * \frac{logK}{n_d} \tag{2.2}$$

onde $freq(t_i, d)$ é igual à frequência com que o termo t ocorre em d.
As principais vantagens do Modelo Vectorial são a sua simplicidade, facilidade de cálculo das similaridades com eficiência, e o facto de permitir trabalhar com conjuntos genéricos de dados. Note-se que existem outros modelos propostos para calcular a similaridade entre documentos (ver [58]). O resultado do cálculo computacional dos índices de similaridade entre uma consulta e todos os documentos do sistema permite ordenar os resultados por ordem decrescente de relevância, isto é, similaridade. Desta modo é possível apresentar ao utilizador em primeiro lugar os documentos que no sistema de recuperação de informação são mais similares com a consulta, e que podem coincidir, ou não, com o resultado esperado pelo utilizador. A relevância é uma medida subjectiva que o sistema tem para determinar dentre os resultados possiveis, os que lhe poderão agradar e sejam adequados ás necessidades informativas do utilizador [64].

2.3.2 Modelo booleano

O Modelo Booleano baseia-se na teoria de Álgebras de Boole e em conceitos de lógica para ordenar os documentos relevantes de uma determinada consulta. Neste modelo, o conjunto de documentos D é representado por um conjunto de palavras-chave, ou seja, por um conjunto de termos Γ. A indexação realiza-se associando um peso binário a cada termo $t \in \Gamma$, isto é, a função de indexação

$$W : D \times \Gamma \rightarrow \{0, 1\} \tag{2.3}$$

que pode ser definida por

$$W(d, t) = \begin{cases} 1, & \text{se } t \text{ aparece pelo menos uma vez em } d \\ 0, & \text{caso contrário.} \end{cases} \tag{2.4}$$

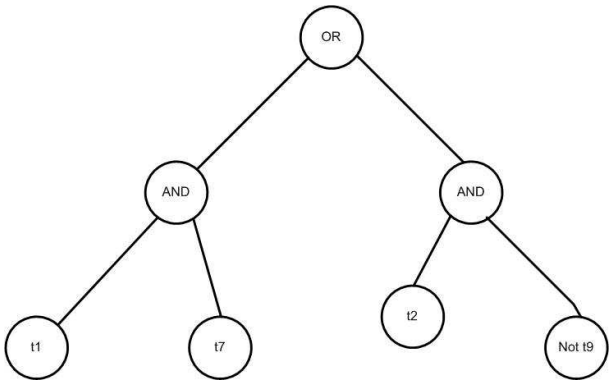


Figura 2.3: Exemplo de uma consulta com o modelo Booleano

onde $t \in \Gamma$ e $d \in D$.

Uma consulta Q consiste em expressões de termos com um ou mais operadores lógicos *AND, OR e NOT*. O grau de similaridade entre o documento e a consulta será um valor binário que resultará do valor lógico das expressões booleanas formadas pelos pesos binários dos termos nos documentos definidos em (2.4). O documento será classificado como relevante se o seu grau de similaridade em relação à consulta é um e classificado como não relevante se o seu grau de similariedade em relação à consulta é igual a zero. Neste modelo as consultas comportam-se então como expressões booleanas que contêm um conjunto de termos Γ e um ou mais operadores booleanos. Um exemplo deste tipo de consulta seria

$$(t_1 \text{ AND } t_7) \text{ OR } (t_2 \text{ AND NOT } t_9)$$

que graficamente pode ser representado como se mostra na figura 2.3. . Neste caso, o grau de similaridade é calculado através do valor lógico da expressão booleana

$$(W(d,t_1) \text{ AND } W(d,t_7)) \text{ OR } (W(d,t_2) \text{ AND NOT } W(d,t_9))$$

que tomará o valor de 1 se os termos forem relevantes para a consulta e 0 caso contrário. Os principais problemas do modelo booleano são a ausência de ordem total na resposta e as respostas podem ser conjuntos vazios ou muito grandes. Já em termos de vantagens a facilidade de implementação, e a expressividade completa das expressões são factores favoráveis na utilização deste modelo [64].

2.3.3 Modelo probabilístico

O Modelo Probabilístico tem como ideia principal de todo o seu processo de implementação o cálculo da probabilidade de determinado documento ser ou não relevante para certa consulta. A ordenação é feita dinamicamente, pesando os termos da consulta relativamente aos documentos baseando-se fundamentalmente no príncipio da ordenação probabílistica [36]. Defina-se então por $D = \{d_1, d_2, ..., d_i, ...\}$ o conjunto de todos os documentos e Q uma consulta.
Antes de se proceder à explicação deste modelo torna-se necessário enunciar o teorema de Bayes.

Teorema 2 *[18] Sejam $A_1, A_2, ..., A_n$ eventos mutuamente exclusivos cuja união é o espaço amostral S, onde $P(A_i) \neq 0$, $i \in \{1, 2, ..., n\}$. Então para qualquer evento B de S tal que $P(B) \neq 0$, tem-se que*

$$P(A_k|B) = \frac{P(A_k B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{N} P(A_i)P(B|A_i)}. \tag{2.5}$$

Assumindo que a relevância de um documento é independente da relevância de todos os outros, um documento d_i é relevante a uma consulta Q quando

$$P(R|d_i) > P(\overline{R}|d_i) \tag{2.6}$$

Assim, dada uma consulta Q o modelo probabilístico atribui a cada documento um peso, $W_{d_i|Q}$ como sendo

$$W_{d_i|Q} = \frac{P(R|d_i)}{P(\overline{R}|d_i)} \tag{2.7}$$

A fórmula (2.7) calcula a probabilidade de um documento ser ou não relevante e assim,

$$W_{d_i|Q} = \frac{P(R|d_i)}{P(\overline{R}|d_i)} \rightarrow \begin{cases} > 1, & \text{se relevante;} \\ < 1, & \text{caso contrário.} \end{cases} \tag{2.8}$$

Aplicando o Teorema de Bayes, (Teorema 2) a (2.7) resulta,

$$W_{d_i|Q} = \frac{P(d_i|R)P(R)}{P(d_i|\overline{R})P(\overline{R})}, \tag{2.9}$$

onde $P(d_i|R)$ representará a probabilidade que, dado um documento relevante para a consulta Q, este seja d_i e $P(d_i|\overline{R})$ representará a probabilidade que, dado um documento não relevante para a consulta Q, este seja d_i. Para calcular $P(d_i|R)$ e $P(d_i|\overline{R})$, o documento d_i pode ser representado por um conjunto de termos $\Gamma = \{t_1, ..., t_n\}$ com determinado peso cada. Estes pesos são valores binários, em que cada termo tem peso 1 ou 0, consoante ser ou não relevante para a consulta. Desta forma resulta,

$$P(d_i|R) = \prod_{k=1}^{n} P(t_k|R) \tag{2.10}$$

e

$$P(d_i|\overline{R}) = \prod_{k=1}^{n} P(t_k|\overline{R}). \tag{2.11}$$

Seja $r_k = P(t_k = 1|R)$ a fórmula (2.10) pode ser reescrita da seguinte forma,

$$P(d_i|R) = \prod_{k=1}^{n} r_k^{t_k}(1 - r_k^{t_k})^{1-t_k} \tag{2.12}$$

onde $r_k^{t_k}$ representa a probabilidade que, dado o documento d_i, este é relevante para a consulta Q, se o termo de ordem k está presente em d_i. De forma análoga, considerando $s_k = P(t_k = 1|\overline{R})$ a fórmula (2.11) pode ser escrita da seguinte forma

$$P(d_i|\overline{R}) = \prod_{k=1}^{n} s_k^{t_k}(1 - s_k^{t_k})^{1-t_k}. \tag{2.13}$$

Substituindo (2.12) e (2.13) em (2.9) tem-se

$$W_{d_i|Q} = \sum_{k=1}^{n} t_k \times w_k + C \tag{2.14}$$

onde $t_k \in \{0, 1\}$,

$$w_k = log\frac{r_k}{1 - r_k} + log\frac{1 - s_k}{s_k} \tag{2.15}$$

e

$$C = log\frac{P(R)}{P(\overline{R})} + \sum_{k=1}^{n} log\frac{1 - r_k}{1 - s_k}. \tag{2.16}$$

Desta forma, observa-se que para avaliar a similaridade entre a consulta Q e o documento d_i basta apenas avaliar os pesos para os termos da consulta. Assim, garante-se que

$$Sim(d_i, Q) \approx W_{d_i|Q}. \tag{2.17}$$

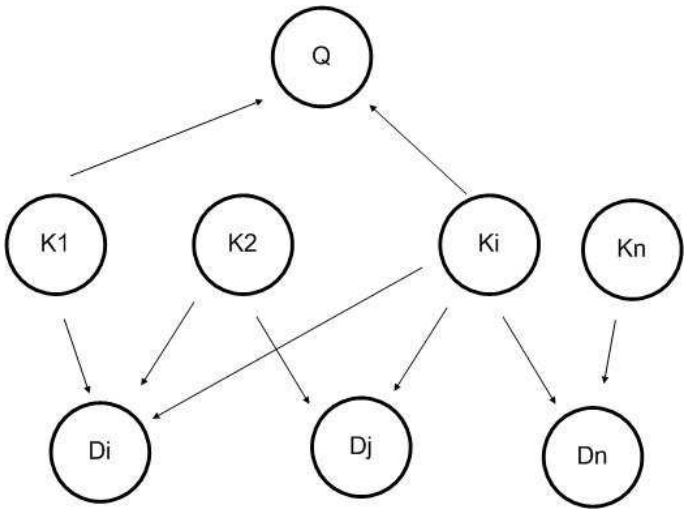


Figura 2.4: Rede bayesiana para uma consulta Q composta por termos K_1 e K_i

Este modelo tem como vantagem, além do bom desempenho prático, o princípio probabilístico de ordenação, que uma vez garantido, resulta num comportamento óptimo. A desvantagem é que este comportamento depende da precisão das estimativas de probabilidade [64]. Para além disso, o método não explora a frequência dos termos no documento e ignora o problema de filtragem de informação.

2.3.4 Redes bayesianas

Nos sistemas tradicionais de recuperação de informação baseados em conteúdo, documentos e consultas são representados como um conjunto de termos. No modelo de Redes Bayesianas a recuperação de informação é também baseada em conteúdo, documentos e consultas são tratados de forma análoga. A figura 2.4 ilustra uma Rede Bayesiana que reflecte esta simetria entre documentos e consultas. Nesta rede, cada nodo D_j modela um documento d_j, um nodo Q modela a consulta q do utilizador e os nodos k_i modelam os termos do conjunto. Neste modelo a ordenação é feita pela similaridade entre o documento d_j e uma consulta q e é dada pela probabilidade

$$P(d_j|q) = \eta \sum_{\forall k} P(d_j|k)P(q|K)P(K), \tag{2.18}$$

onde η é uma constante de normalização e $P(d_j|k)$ é definida pela regra da probabilidade total e as independências modeladas na rede. Detalhes da derivação desta equação podem ser encontradas em [49]. Também em [49], é proposto um Modelo Bayesiano que combina as informações de apontadores e conteúdo da rede telemática.

Os modelos descritos anteriormente são exemplos de modelos que fornecem uma ordenação de documentos baseado no conteúdo. Mas apresentar um ordenação de documentos baseado na informação sobre as estruturas de apontadores obtidos para cada documento e conjugar essa informação com o conteúdo são estratégias que são cada vez mais úteis e mais aplicadas nos SVRI. A exploração deste assunto é realizada em detalhe no próximo capítulo.

Capítulo 3

Algoritmos de Pesquisa Baseados em Apontadores

3.1 Introdução

A relevância da informação num conjunto de dados e o modo como esta será pesquisada é um dos principais requisitos que os Sistemas de Recuperação de Informação (IR) terão de ter. Seja um requisito da perspectiva do sistema seja da perspectiva do utilizador. Hoje em dia, o primeiro e grande conjunto de documentos a ser pesquisado é o da Internet [53]. Criar um sistema de armazenamento de informação que não fosse controlado nem controlável foi a ideia subjacente à criação da Internet. No entanto, essa ideia tem-se vindo a desmoronar. Diariamente são apresentados aos utilizadores aproximadamente 15 milhões de páginas das quais os utilizadores leêm cerca de 100 páginas, e.g. através do Google.

A organização e normalização da Internet é feita por vários intervenientes. No entanto, mesmo que alguns deles, por exemplo administradores da rede, estabeleçam regras e princípios de modo a estruturar a rede estas nem sempre são obedecidas. Por exemplo, apesar dos browsers actualmente existentes na Internet serem estáveis e robustos,

com detectores e correctores automáticos de erros continua a existir um número de páginas escritas em html que sintacticamente contêm inúmeros erros. Outros intervenientes como é o caso de, instituições estatais e internacionais, por exemplo a ONU, desenvolvem regras para estabelecer uma norma única, "ISSS - Information Society Standardization System". As universidades e investigadores são intervenientes que frequentemente propõem trabalhos de investigação com o objectivo de criar um sistema universal e indexado de dados.

Deste modo, a catalogação da informação inserida na rede telemática é uma preocupação cada vez mais frequente, mas a velocidade com que aumenta a quantidade de informação na rede torna cada vez mais difícil estabelecer um método eficaz de recuperar e apresentar essa informação. Assim, as tentativas de organização e estruturação do actual excesso de informação presente na rede telemática são um problema ainda em aberto. Pode-se afirmar que é o factor mais visível nos trabalhos de investigação efectuados em "Information Retrievel" (IR), apesar da investigação de um sistema ideal que garanta resultados mais relevantes ainda não tenha sido alcançado. De facto, na documentação científica encontram-se diversos estudos sobre sistemas de pesquisa, a sua eficácia, características e modos de funcionamento, mas de um modo geral, apresentam-se como um sistema de recolha de informação.

A criação de um sistema de recuperação de informação engloba quatro etapas fundamentais: um processador de documentos, um processador de perguntas, uma função de pesquisa e comparação e a possibilidade de ordenar hierarquicamente os documentos de modo a serem apresentados de forma ordenada. Cada sistema de pesquisa contém particularidades resultantes de diferentes filosofias e procedimentos no seu desenvolvimento. Todos os dias aparecem e desaparecem servidores e portais ou então fundem-se entre si.

Usando diferentes sistemas de pesquisa para pesquisar informação tendo por base os mesmo termos ou conceitos, tanto se pode obter respostas substancialmente diversas, como se podem reconhecer respostas já apontadas por outros sistemas de pesquisa. Por essas razões surgem os meta motores de pesquisa cuja pesquisa é dirigida por um

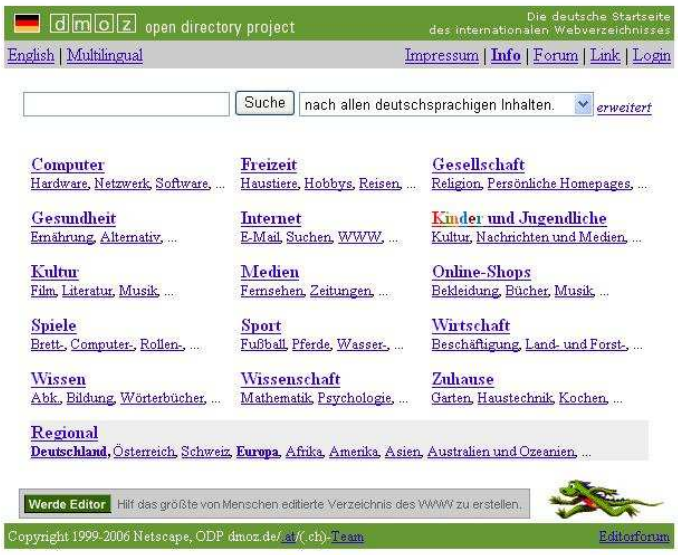


Figura 3.1: Site DMOZ

conjunto ampliado de sistemas de pesquisa.
Genericamente, os sistemas de pesquisa podem ser enquadrados em duas classes consoante o seu funcionamento:

- Sistemas manipulados manualmente, onde a manipulação é feita pelos administradores do sistema que catalogam e verificam o conteúdo das páginas. Exemplos são o DMOZ (ver figura 3.1), ZEAL, etc., mas o mais popular continua a ser o Yahoo.

- Sistemas manipulados por *Crawler/Spiders/Robots*, que como o próprio nome indica são manipulados por um programa *Spider ou Crawler* [23] que percorre de forma constante e automática a Internet, passando página a página, através da sua estrutura de apontadores, catalogando e armazenado as páginas encontradas ao longo do percurso. Exemplos desses sistemas são o Google, (ver figura 3.2).

Como se pode depreender, a recuperação e pesquisa de informação na rede telemática é muito diferente da que pode ser feita em bases de dados tradicionais, mas é a partir do conceito de indexação de bases de dados tradicionais que se inspiram todos os estudos

Figura 3.2: Site Google

sobre sistemas de recuperação de informação da rede telemática.

Desde meados de 1995, que se procura encontrar os tipos de metadados certos para correctamente identificar um documento na rede telemática. Nesse mesmo ano, num workshop em Dublin [55], estabeleceram-se uma série de elementos identificadores considerados essenciais para catalogação dos documentos. No entanto, as categorias propostas não satisfazem na totalidade as necessidades comerciais e de utilização da rede telemática.

O desenvolvimento de sistemas de pesquisa de informação tendo em conta a perspectiva do utilizador surgem na década de 70, os seus pioneiros foram Saraceviv e Kantor, que tiveram como principal objectivo de investigação identificar o comportamento e a satisfação do utilizador para posterior enquadramento na efectividade do sistema. A partir dai, os sistemas evoluem no sentido de incluir neles o comportamento humano, englobando neles processos de filtragem, aglomeração, utilidades e muitas outras funcionalidades. Num dos poucos artigos existentes sobre as funções dos sistemas de recolha de informação da perspectiva do utilizador, Thorsten Joachims [24] propõe

vários testes para determinar o sucesso de recuperação de informação centrada no utilizador, muitos deles com deficiência de raciocínio, pois misturam dados de sistemas de pesquisa distintos como é o caso do Google com o MSNSearch.

Mais do que desenvolver novos algoritmos de ordenação a comunidade científica em IR tem direccionado a sua atenção para o melhoramento dos algoritmos já existentes. No âmbito deste trabalho, nomeadamente no estudo das lacunas e das potencialidades dos actuais processos de classificação e organização de dados, fica claro que os actuais algoritmos devem continuar a ser considerados, embora se devam acrescentar novas abordagens: o entendimento dos documentos, a ajuda ao utilizador no âmbito da definição da consulta e a organização do conjunto de resultados.

Depois de se ter apresentado brevemente o estado da arte no domínio dos sistemas de pesquisa, nas próximas secções apresentam-se algoritmos de pesquisa mais utilizados e que se baseiam em estrutura de apontadores.

No capítulo anterior, considerou-se que cada página existente na rede telemática é caracterizada e diferenciada pelo seu conteúdo e foram explorados alguns dos modelos tradicionais de recuperação de informação através do conteúdo das páginas [3].

Contudo, outro tipo de informação pode ser extraída da rede telemática para posterior utilização na recuperação da informação, o conhecimento da estrutura de apontadores da rede. Com este tipo de informação foram criados, novos algoritmos de pesquisa de informação, usando a estrutura de apontadores da rede telemática.

De seguida, são apresentados alguns conceitos e resultados importantes para a explanação do presente capítulo.

Definição 3 *Um grafo G, é um par de conjuntos (V, E), onde $V = \{v_1, v_2, ..., v_n\}$ é o conjunto dos nodos de G e $E = \{e_1, e_2, ..., e_m\}$ é o conjunto dos arcos (orientados) de G, em que a cada um deles corresponde um subconjunto de V com cardinalidade igual a 2, isto é, $e_k = \{v_{k_i}, v_{k_j}\}$ para cada $k \in \{1, ..., m\}$.*

Definição 4 *Um grafo G diz-se* conexo *se não admite qualquer partição para além da*

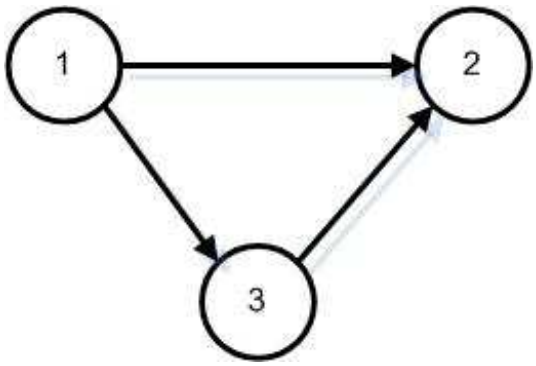


Figura 3.3: Grafo representativo de uma rede telemática

trivial, isto é, $G = G \cup \emptyset$.

Definição 5 *Dado um grafo G conexo, um nodo i de G diz-se uma* autoridade *se existem nodos com arcos a apontarem para i.*

Definição 6 *Dado um grafo G conexo, um nodo i de G diz-se um* hub *se dele sairem arcos a apontarem para outros nodos de G.*

Definição 7 *A matriz $A = [a_{ij}]$, $i, j \in 1, ..., n$, diz-se a matriz de* adjacência *associado ao grafo $G = (V, E)$, com $V = \{v_1, ..., v_n\}$ e é tal que*

$$a_{ij} = \begin{cases} 1, & \textit{existe um nodo de } i \textit{ para } j; \\ 0, & \textit{caso contrário.} \end{cases}$$

A rede telemática pode ser representada por um grafo (orientado) $G = (V, E)$ (ver definição 3), onde cada página da rede é representada por um nodo do grafo e as ligações (apontadores) entre as páginas pelos arcos desse grafo.
Na figura 3.3 está um grafo que serve de exemplo para uma rede telemática. As páginas podem ser vista como autoridades (ver definição 5) ou como hub (ver definição 6).
A figura 3.4 mostra dois exemplos concretos do que pode ser uma autoridade ou um hub. A popularidade de certa página pode ser indirectamente medida através do grau de autoridade que esta representa na estrutura, i.e., quanto mais páginas uma página tiver a apontar para si mais popular se torna essa página.

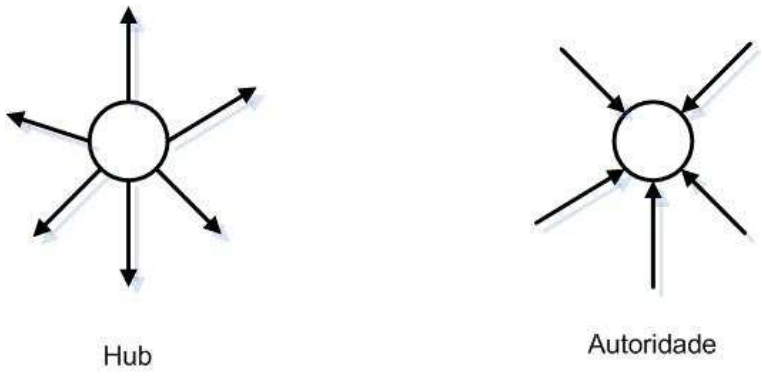


Figura 3.4: Autoridades e Hubs de uma rede telemática

Os três algoritmos predominantes de extracção de informação na rede telemática a partir da estrutura de apontadores são o PageRank (utilizado no método de pesquisa feita pelo Google) o HITS e o SALSA, sendo este último uma combinação dos dois anteriores.

3.2 Algoritmo Hits

O ano de 1998 foi um ano importante para a investigação de modelos de análise de apontadores. Um jovem cientista, Jon Kleinberg, professor assistente na Universidade da Cornell [3] trabalhava no projecto de pesquisa de informação designada por Hits. O algoritmo foi desenvolvido tendo por base a ideia inovadora de tirar partido da estrutura de apontadores na rede telemática. Esta ideia foi apresentada no *9º Simpósio Anual ACM-SIAM*. Desde então, o algoritmo faz parte do projecto de pesquisa *CLEVER da IBM Almaden Research Center* [4]. Ideias similares à deste algoritmo são também usadas pelo sistema de pesquisa Theoma, desenvolvido por Ask Jeeves [54].
Em particular, este algoritmo é muito utilizado em pesquisas bibliométricas. A bibliometria é o estudo da escrita de documentos e da sua estrutura de citações. Tais pesquisas utilizam a estrutura de citação do corpo de um documento para produzir uma maneira numérica de medir a importância e o impacto dos documentos [12].

Observe-se que, o algoritmo Hits não é aplicado ao grafo representativo de toda a estrutura da rede telemática, mas sim a (sub)grafos desta estrutura, usualmente formados por 1000 a 5000 nodos, resultantes da pesquisa tradicional de termos numa consulta feita pelo utilizador.

A ideia fundamental deste algoritmo é a de que páginas que são bons hubs apontam para boas autoridades e boas autoridades são apontadas por bons hubs [35]. Para cada página da rede telemática pode-se calcular o peso de hub e o peso de autoridade.

Considere-se uma página i com peso de autoridade x_i e com peso de hub y_i. Seja E o conjunto de todos os nodos do (sub)grafo que representa a rede telemática (ou o universo de pesquisa) e e_{ij} o arco que representa o apontador que vai de i para j. Sejam $H(i) = \{j : \exists\ e_{ij} \in E\}$ e $A(j) = \{i : \exists\ e_{ij} \in E\}$, os conjuntos de todos os hubs e autoridades de G, respectivamente.

Assumindo que a cada página, é-lhe atribuido um peso inicial de autoridade $x_i(0)$ e um peso inicial de hub $y_i(0)$, usualmente vectores com entradas todas iguais a um, o algoritmo Hits actualiza de forma iteractiva os pesos de hubs e autoridades através da seguinte fórmula

$$x_i(k) = \sum_{j \in A(i)} y_j(k-1) \quad e \quad y_i(k) = \sum_{j \in H(j)} x_j(k) \tag{3.1}$$

com $k \in \mathbb{N}$.

Estas equações podem ser definidas na forma matricial, recorrendo à matriz de adjacência A (ver definição 7), do grafo que representa a rede telemátca. Na forma matricial as equações 3.1 apresentam-se por

$$X^k = A^T Y^{k-1} \quad e \quad Y^k = AX^k. \tag{3.2}$$

Este processo permitirá encontrar dois subconjuntos de páginas, um de hubs e outro de autoridades, ordenados de acordo com os seus pesos de hubs e autoridades, respectivamente.

3.2.1 Implementação do algoritmo Hits

Suponha-se que um determinado utilizador pretende fazer uma pesquisa sobre um assunto inserindo um conjunto de palavras-chave no sistema, produzindo uma consulta com os termos a pesquisar. O resultado para essa consulta será um conjunto finito de páginas que contêm os termos da pesquisa. Para além disso, essas páginas contêm apontadores entre elas. A popularidade entre estas pode ser calculada através do algoritmo Hits e apresentadas ao utilizador ordenadas por popularidade. Este conjunto de páginas designa-se por *universo da consulta.*

A implementação do algoritmo para calcular a ordenação de hubs e autoridades das páginas envolve dois passos principais: um primeiro passo, consiste na construção do grafo que represente as ligações e as páginas envolvidas na consulta que foi feita; e o segundo passo, é calcular para cada página o seu peso de autoridade e de hubs através das fórmulas (3.1). Para calcular os pesos de autoridades e hubs das páginas de forma eficiente, recorre-se a um método designado por "Power Method" [63], que se resume à determinação dos valores próprios (ver definição 8) das matrizes AA^T e A^TA para posterior apresentação dos valores de ordenação de hubs e autoridades, respectivamente.

Definição 8 *Seja E um espaço vectorial sobre um corpo ß ($\Re$ ou C) e $T : E \to E$ em endomorfismo de E. Um vector não nulo $\overrightarrow{v}$ de E diz-se um vector próprio de T se existe um escalar $\lambda \in$ ß tal que:*

$$T(\overrightarrow{v}) = \lambda \overrightarrow{v}$$

O escalar λ diz-se valor próprio de T associado ao vector $\overrightarrow{v}$. Ao conjunto dos valores próprios de T dá-se o nome de espectro de T e designa-se por $\sigma(T)$.

Os custos computacionais do cálculo dos valores próprios podem ser reduzidos, sendo benéfico para a implementação do algoritmo. De forma resumida, tem-se que o algoritmo Hits é um algoritmo iterativo que envolve apenas o cálculo dos valores próprios

associados às matrizes A^TA e AA^T, como se pode observar seguidamente.

Algoritmo Hits: Cálculo dos vectores autoridade e hubs

Entrada: $Y^{(0)} = \mathbb{1}$, onde $\mathbb{1}$ é uma vector coluna com entradas todas iguais a um.

1: **while** (não existe convergência) **do**

2: $X^k = A^T Y^{k-1}$

3: $Y^k = AX^k$

4: $k = k + 1$ e normalizar X^k e Y^k (ver secção 3.2.2)

5: **end while**

Saida: X^k e Y^k

Os passos 2 e 3 do algoritmo [35], podem ser simplificados através da substituição das equações por

$$\begin{cases} X^k = A^T A X^{k-1} \\ Y^k = AA^T Y^{k-1} \end{cases}$$

A aplicação deste algoritmo garante sempre a existência dos vectores hubs e autoridades das páginas como se verifica pelo seguinte resultado:

Teorema 9 *[2] Seja G o grafo orientado com matriz de adjacência A. Os vectores autoridades e hubs dados pelo algoritmo Hits existem e têm entradas todas não negativas.*

Prova Uma matriz quadrada simétrica diz-se não negativa se e só se todos os seus valores próprios forem não negativos. Como as matrizes AA^T e A^TA são matrizes quadradas, simétricas com valores próprios todos não negativos então pode-se garantir que elas são não negativas. O espectro de AA^T é igual ao de A^TA, i.e. $\sigma(AA^T) = \sigma(AA^T)$, então considere-se $M = AA^T$. Como qualquer matriz não negativa possui sempre um valor próprio positivo $\overline{\lambda}$, tal que em módulo nenhum dos outros valores próprios associados o excede, isto é, $\lambda \leq |\overline{\lambda}|$ $\forall \lambda \in \sigma(M)$, então pode-se escolher o vector próprio associado ao valor próprio $\overline{\lambda}$, ficando assim garantida que qualquer entrada do vector próprio é não negativa. □

3.2.2 Convergência do algoritmo

O algoritmo Hits é um algoritmo iterativo que em termos computacionais envolve apenas o cálculo dos vectores Hits por aplicação do "Power Method" às matrizes A^TA e AA^T. Examine-se a convergência do algoritmo pelas propriedades de convergência do "Power Method".

Seja $B \in M_{n\times n}(\mathbb{R})$ uma matriz diagonalizável, formada por n vectores próprios independentes, $\{u_1, ..., u_n\}$, associados a n valores próprios distintos $\{\lambda_1, \lambda_2, ..., \lambda_n\}$, tais que $\mid \lambda_1 \mid > \mid \lambda_2 \mid \geq \mid \lambda_3 \mid \geq ... \geq \mid \lambda_n \mid$. Este vectores próprios formam uma base em $\mathbb{R}^n$.

Considerando que o vector de autoridades inicial $X^{(0)}$ pode ser descrito por

$$X^{(0)} = \alpha_1 u_1 + \alpha_2 u_2 + ... + \alpha_n u_n \tag{3.3}$$

onde $\alpha_1, ..., \alpha_n$ são escalares e multiplicando ambos os membros da equação (3.3) por B^k resulta

$$B^k X^{(0)} = B^k(\alpha_1 u_1 + \alpha_2 u_2 + ... + \alpha_n u_n) = \alpha_1 \lambda_1^k \left(u_1 + \sum_{j=2}^{n} \frac{\alpha_j}{\alpha_1} (\frac{\lambda_j}{\lambda_1})^k u_j \right) \tag{3.4}$$

Pela monotonia dos valores próprios, pode-se afirmar que $\lambda_1 = \rho(B)$ (ver definição 10). Neste caso, $(\frac{\lambda_j}{\lambda_1})^k \to 0$ quando $k \to \infty$ e se $\alpha_1 \neq 0$ então $B^k X^{(0)} \to \alpha_1 \lambda_1^k u_1$. O "Power Method" normaliza os produtos $BX^{(k-1)}$ para evitar "overflow" ou "underflow", por esta razão estes convergem para u_1. De facto, o algoritmo Hits converge se $\lambda_1 = \rho(B)$ e se $X^{(0)}$ tem uma componente na direcção do vector próprio associado a λ_1, neste caso u_1. Na prática, a convergência depende de $\frac{|\lambda_2|}{|\lambda_1|}$, pois esta razão dita a ordem de convergência do algoritmo.

Assim sendo, se o algoritmo converge para o vector próprio associado a $\rho(B)$ depois de k iteracções então $[X^{(k)}]^T B X^{(k)} \approx [X^{(k)}]^T \lambda_1 X^{(k)} = \lambda_1 \|X^{(k)}\|^2 = \lambda_1$, pois $\|X^{(k)}\|^2 = 1$ devido ao facto de se ir normalizando $X^{(k)}$ a cada iteracção.

No entanto, poderão surgir problemas para garantir a unicidade dos valores de autoridades e de hubs dos documentos. Enquanto se verificar $\lambda_1 > \lambda_2$, pode acontecer que λ_1 se repita como raiz do polinómio característico associado a B o que por outras palavras quererá dizer que diferentes limites de vectores de autoridade (hubs) poderão produzir diferentes escolhas do vector inicial para o processo computacional.

Definição 10 *O raio espectral $\rho(A)$ de uma matriz quadrada A é definido como o máximo das amplitudes dos valores próprios de A.*

Um exemplo simples retirado de [2] poderá demonstrar com facilidade o problema inerente. Nomeadamente, sejam

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

e

$$A^T A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

A matriz de autoridade $A^T A$ (e a matriz de hubs AA^T) têm dois valores próprios distintos $\lambda_1 = 2$ e $\lambda_2 = 0$, com multiplicidade dois. Para $X^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})^T$, o método descrito converge para $X^{(\infty)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)^T$. No entanto, para $X^{(0)} = (\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2})^T$, o método descrito converge para $X^{(\infty)} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)^T$ como se pretendia mostrar.

J. Miller et al. [3] mostraram que o algoritmo Hits é "mal-condicionado", em algumas estruturas da rede telemática, o que permite afirmar que:

i) Este algoritmo poderá retornar, em alguns casos, um vector de ordenação que não é único e dependente do vector inicial considerado;

ii) O algoritmo poderá devolver uma ordenação de vectores que têm entradas com pesos de hubs e autoridades inapropriados ou até mesmo nulos para certas estruturas da rede telemática.

J. Miller et al. [3] mostraram ainda que o mal-condicionamento esta relacionado com a conectividade de G. Foram recentemente propostas modificações ao algoritmo Hits, denominado Exponencial Input para HITS [3], por alteração da matriz de adjacência para uma matriz exponencial, por forma a garantir a unicidade dos vectores próprios, sem entradas com pesos inapropriados ou pesos nulos, desde que a estrutura a considerar seja conexa.

Uma das grandes vantagens deste algoritmo é o de devolver uma ordenação dupla, isto é,

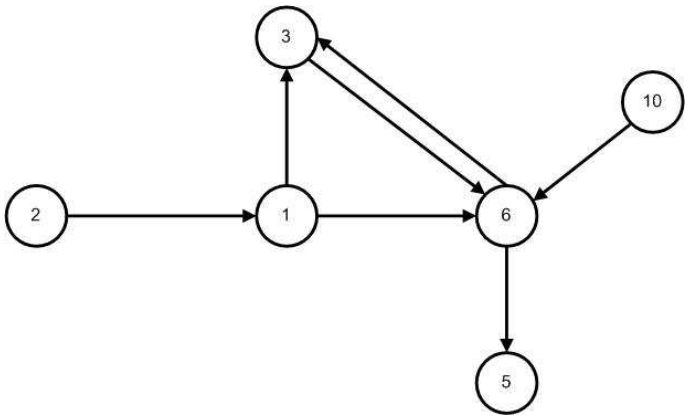


Figura 3.5: G: grafo de vizinhança dos documentos v_1 e v_6

poderão ser apresentadas duas listas de documentos ordenadas: uma lista com todos os documentos com maior comportamento do tipo autoridades, relativamente a uma consulta, e outra com os documentos mais do tipo Hub. Um utilizador poderá estar mais interessado numa ou noutra lista dependendo do tipo de aplicação e pesquisa. O algoritmo Hits também consegue reduzir o problema geral da pesquisa na rede telemática a um problema menor ao manipular matrizes de pequena dimensão o que facilita no cálculo dos vectores próprios associados a essas matrizes.

3.2.3 Exemplo

Seguidamente apresenta-se um pequeno exemplo para demonstrar a implementação do algoritmo Hits. Em primeiro lugar, um determinado utilizador apresenta uma consulta de termos para pesquisa ao sistema. Suponha-se que o subconjunto de nodos que contêm os termos da consulta é $V_Q = \{v_1\ , v_6\}$, isto é, os documentos 1 e 6 contêm termos da consulta do universo $V = \{v_1, v_2, v_3, v_4, v_5,\ v_6\}$. Contrua-se um sub grafo da estrutura da rede telemática que contém todos os vizinhos de v_1 e de v_6 e as respectivas ligações entre eles, onde tal operação conduz a um grafo G (figura 3.5). Para G, a matriz de adjacência A é dada por

$$\mathrm{A}=\begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

As matrizes autoridade e hubs são, respectivamente,

$$M_A = A^T A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ e } M_H = AA^T = \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

A partir da execução do algoritmo Hits descrito, envolvendo o cálculo dos valores e vectores próprios associados, resultam os seguintes resultados para as autoridades, X, e para os hubs, Y, após k iteracções,

$$X^T = (0,\ \ 0,\ \ 0.3660,\ \ 0.1340,\ \ 0.5,\ \ 0),$$

$$Y^T = (0.36660,\ \ 0,\ \ 0.2113,\ \ 0,\ \ 0.2113,\ \ 0.2113).$$

De onde se conclui que, o documento v_6 é o documento mais do tipo autoridade para a consulta, enquanto que o documento v_1 é o melhor hub para esta consulta.

3.3 Algoritmo PageRank

Originalmente descrito por *Larry Page e Sergey Brin*, fundadores do motor de pesquisa Google, o algoritmo PageRank é o coração do Google e define-se como um algoritmo que ordena as páginas da rede telemática pela sua relevância (popularidade) e foi inspirado no "Science Citation Index". Enquanto estudantes da Universidade de Stanford, *Larry Page e Sergey Brin* [56] pensavam na criação de um sistema para organizar teses de mestrado

e doutoramento, acabaram por concluir que uma obra académica ganhava em termos de relevância quando citava obras do mesmo assunto. Para além disso, uma obra citada por outras também ganhava importância.

Existem dois modos distintos de interpretar o PageRank, sendo um destes no âmbito de cadeias de Markov. O PageRank de uma página p é a probabilidade de visitar p através de um caminho aleatóriamente escolhido da rede telemática, onde o conjunto de estados do caminho é o conjunto de páginas. Cada passo do método iteractivo é feito mediante dois tipos de escolha: visitar p através de uma página escolhida aleatóriamente ou através de uma página q escolher aleatóriamente a próxima página do caminho aleatório que consiga ir de q até p. A outra interpretação do PageRank é baseada na ideia de que uma determinada página da rede telemática é importante se outra página também importante tem um apontador para ela, vista como uma maneira de calcular a importância duma página através dum voto nela. Cada apontador é um voto e aumenta a importância de determinada página, se se considerar recursivamente a estrutura de apontadores. Neste contexto, o PageRank pode ser interpretado como um contador "on-line" de votos, onde umas páginas votam para importância de outras, e o resultado da votação é apresentado pelo valor de PageRank de cada página e reflectido nos resultados da pesquisa [56]. No entanto, nem todos os votos têm a mesma relevância, por exemplo, um voto duma página com pouca importância pouco altera a importância duma página com muita importância.

Estas características podem ser traduzidas através da definição do índice de ordenação $\tau(p)$, de determinada página p,

$$\tau(p) = \sum_{Q \in B_p} \frac{\tau(Q)}{|Q|}, \quad onde \quad B_p = \{todas\ as\ paginas\ que\ apontam\ para\ p\} \quad , \tag{3.5}$$

|Q|=número de apontadores que saem de Q .

Esta definição para o índice de ordenação de uma página é recursiva, dai a necessidade de computacionalmente existir iteracção. Considerem-se n páginas, $p_1, p_2, ..., p_n$ com determinado índice de ordenação inicial, arbitrariamente escolhido, $r_0(p_i) = \frac{1}{n}$. Então, sucessivamente, pode-se refinar a cada iteracção o índice de ordenação das páginas pelo cálculo computacional de

$$r_j(p_i) = \sum_{Q \in B_{p_i}} \frac{r_{j-1}(Q)}{|Q|} \quad para\ j \in \mathbb{N}, \tag{3.6}$$

acompanhado do cálculo final do vector $\pi_j^T = (r_j(p_1), r_j(p_2), ..., r_j(p_n))$, ou de forma iteractiva do cálculo computacional de

$$\pi_j^T = \pi_{j-1}^T P \ , \tag{3.7}$$

onde P é a matriz com $p_{ij} = \frac{1}{|p_i|}$ se existe um apontador de P_i para P_j e 0 caso contrário.
O vector PageRank é definido, caso exista o limite, por

$$\pi^T = lim_{j\to\infty}\pi_j^T \tag{3.8}$$

onde a componente π_i caracteriza o PageRank da página p_i.
De forma abstracta, o anterior processo é a ideia geral do cálculo do PageRank para uma rede telemática, no entanto, por razões teóricas e práticas, essencialmente devido à necessidade de garantia de convergência do algoritmo, a matriz P tem de ser ajustada. A matriz P, utilizada no Google, é não negativa com a soma das linhas todas iguais a um ou a zero. Soma de linhas iguais a zero corresponde a páginas que não têm saída de apontadores para outras páginas sendo os nodos correspondentes designados por nodos "dangling" do grafo. Por agora, assume-se que não existem nodos "dangling" ou que este problema é resolvido através da adição de apontadores apropriados para fazer com que a soma das linhas seja sempre igual a um e assim P seja uma matriz estocástica (ver definição 12).

Definição 11 *Uma* Cadeia de Markov *é um processo de Markov cujo espaço de estados é finito ou contável e cujo conjunto de indices é* $T = \{0, 1, 2, ...\}$.

Definição 12 *Uma matriz* $P_{n\times n}$ *diz-se uma matriz estocástica se todos os seus elementos têm valor compreendido entre 0 e 1, isto é,*

$$0 < P_{ij} < 1 \ \forall \ i, j \in \{1, ..., n\},$$

além disso, a soma de todos os elementos de cada linha tem de ser igual a um.

Definição 13 *Uma matriz* I *associada a um grafo orientado* G *diz-se irredutível se todos os nodos de* G *comunicam entre si, i.e., o grafo é conexo.*

Pode-se afirmar, considerando as definições anteriores, que a equação (3.7) representa a evolução de uma Cadeia de Markov (ver definição 11), mais precisamente, esta Cadeia de

Markov é um caminho aleatório definido no grafo da rede telemática.

De facto, a existência de nodos "dangling" pode ser facilmente resolvida através da substituição de cada linha nula de P por $\frac{\mathbb{1}^T}{n}$, onde n define a dimensão da matriz P. Resultando assim uma nova matriz $\overline{P}$, sendo esta estocástica.

Refira-se que, no entanto, a matriz não é necessariamente uma matriz irredutível (ver definição 13), isto é, pode acontecer que nem todos os nodos de G comuniquem entre si, logo outro passo importante é garantir a irredutíbilidade de $\overline{P}$.

Larry Page e Sergey Brin para garantirem a irredutíbilidade da matriz aplicaram o seguinte processo

$$\overline{\overline{P}} = \alpha\overline{P} + (1-\alpha)\mathbb{1}\mathbb{1}^T\frac{1}{n}, \qquad \alpha \in]0,1[\ , \tag{3.9}$$

onde n o número de páginas da rede telemática.

A matriz $\overline{\overline{P}}$ é estocástica (ver definiçao 12), irredutível e conduz a um único vector próprio chamado vector PageRank definido por

$$\pi\overline{\overline{P}} = \pi \quad e \quad \pi\mathbb{1} = \mathbb{1}\ , \tag{3.10}$$

onde $\mathbb{1}$ representa uma coluna de vectores de uns.

Teorema 14 *[3] Seja G um grafo orientado com matriz de adjacência A. Se $\alpha \in]0,1[$ é como na equação (3.9) e $\overline{P}$ é uma matriz estocástica, então o vector PageRank existe e é único. Para além disso, todas as entradas do vector são positivas.*

Prova Se $\alpha > 0$ e $\overline{P}$ uma matriz positiva então a matriz $\overline{\overline{P}}$ é positiva, e desde que $\overline{\overline{P}}^T$ tenha um valor próprio associado simples e positivo com um único vector próprio associado e com entradas todas positivas então pelas propriedades das Cadeias de Markov a convergência para este vector próprio fica provada. □

Este algoritmo é bastante eficiente, mas com o crescimento da quantidade de informação existente na rede telemática este algoritmo é aplicado apenas a subconjuntos de páginas da rede telemática e numa forma adaptada à computação paralela. Observe-se que a actualização do PageRank à medida que são constantemente adicionadas novas páginas envolve custos computacionais muito elevados.

3.3.1 Existência e convergência

O algoritmo PageRank provou ser bastante eficiente para determinar o indíce de ordenação de páginas mais visitadas da rede telemática, podendo ser sempre garantida a existência de um vector de ordenação como se observa pelo seguinte resultado.

Teorema 15 *[5] Para qualquer página $i \in \mathbb{N}$, é sempre garantido que*

$$min_{1 \leq i \leq n}\{\pi_i\} \leq \frac{(1-\alpha)}{n}.$$

Prova De facto, uma página i tem índice de ordenação mínimo se nenhuma outra página refere esta página. Então, neste caso, tem-se

$$\pi_i = [\pi\overline{\overline{P}}]_i = (1-\alpha)(1-\tfrac{1}{n})\pi\ \mathbb{1} = (1-\alpha)(\tfrac{1}{n})$$

o que prova a afirmação realizada. □

Garantida a existência desse vector poderão existir ainda problemas na convergência deste algoritmo.

Teorema 16 *Se uma matriz não negativa é estocástica então tem um valor próprio associado igual a 1 com vector próprio dado por $e = [1,\ 1, ..., 1]^T$. Para além disso, o raio espectral da matriz é igual a 1.*

Seja $\overline{P}$ a matriz estocástica definida a partir da matriz P através da substituição de cada linha nula de P por $\frac{\mathbb{1}^T}{n}$, com n a dimensão de P. Por aplicação do teorema 16 e dado que $\overline{P}$ é uma matriz não negativa e estocástica, então o máximo das amplitudes do valores próprios associados a $\overline{P}$ é igual a 1, isto é, $\rho(\overline{P}) = 1$. Se esta matriz estocástica for redutível, podem-se obter inúmeros valores próprios no circulo unitário, causando problemas de convergência do algoritmo.

Este problema foi identificado pelos inventores do PageRank a quando da elaboração de testes experimentais: um nodo do grafo que não contenha nenhum apontador a sair dele, mas que vai acumulando a cada iteração cada vez mais PageRank poderá causar problemas de convergência. No entanto, forçando a irredutibílidade de $\overline{P}$ os problemas de convergência do algoritmo são contornados. Segue-se o seguinte resultado que vai de encontro ao que se afirmou sobre a convergência do algoritmo e que foi definido por A. Langville [35].

Teorema 17 *Seja A uma matriz estocástica irredutível, então A possui um único valor próprio, $\overline{\lambda}$, no círculo unitário e é tal que:*

$$|\lambda| \leq |\overline{\lambda}|, \forall \lambda \in \sigma(A)$$

para além disso, $\rho(A) = 1$

Com este resultado garante-se a convergência do algoritmo para uma matriz estocástica irredutível. Dessa convergência resulta o vector PageRank π^T associado à matriz estocástica irredutível.

3.3.2 Número de iteracções

Aplicando o método do PageRank a uma matriz estocástica e irredutível P, sabe-se que esta converge para o único vector π de PageRank, sendo a ordem de convergência deste algoritmo importante no processo de obtenção do vector π. Especialmente se se considerar a grande quantidade de operações de multiplicação que são necessárias, em casos reais, estas são da ordem dos biliões.

O número de iteracções do PageRank varia consoante o factor de perturbação α, que se associa a $\mathbb{1}\mathbb{1}^T$ na equação (3.9), para garantir a irredutibilidade de P. Se a rede telemática for irredutível, pode-se concluir que a ordem de convergência do algoritmo aplicada a P é uma ordem tal que $\alpha^k \to 0$, com k o número de iteracções, explicando-se asssim o facto do Google utilizar um valor para $\alpha = 0.85$.

Considerando que o número de iterações necessárias para garantir a convergência do algoritmo com um nível residual de tolerância τ é dado pela expressão

$$\frac{log(10)}{\tau}\frac{log(10)}{\alpha} \quad , \tag{3.11}$$

para $\tau = 10^{-6}$ obtem-se um valor de $\alpha = 0.85$. Portanto, são necessárias aproximadamente 85 iterações para garantir a convergência do vector PageRank. Para $\tau = 10^{-8}$ serão necessárias 114 iterações, para $\tau = 10^{-10}$ serão necessárias 142 iterações e assim sucessivamente. Este factor levou a que os inventores do PageRank assumissem usar entre 50 a 100 iterações, fazendo variar o nível de tolerância entre 10^{-3} e 10^{-7}. Explica-se assim o facto de o Google

usar o valor $\alpha = 0.85$. No entanto, fica-se com a percepção que o Google pode alterar a ordem de convergência do seu algoritmo, fazendo-o consoante a escolha do valor de α, como afirma A. LangVille [34].

3.3.3 Critérios de convergência

O facto do método de cálculo do PageRank descrito ser um método que engloba um grande conjunto de problemas de resolução computacional (cálculo do vector estacionário da cadeia de Markov) da matriz de transição que apresenta um tamanho gigantesco e que em nada facilita na eficiência computacional do algoritmo, fez com que um número elevado de investigadores estudassem e propusessem alternativas à aproximação do vector de ordenação.
Alguns investigadores sugeriram o uso de simples contagem de apontadores que apontam para uma página para fazer o cálculo do PageRank de páginas. Note-se que a ideia original do PageRank é que a importância não advém da quantidade de apontadores para uma página, mas sim, pelo contrário, da qualidade desses apontadores.
Um método foi proposto para acelerar a convergência do algoritmo de PageRank, e fazer com que em cada iteração o esforço computacional diminuísse. Este método foi proposto por Kamvar et al. [52] e é designado de método adaptativo de PageRank. Este método adaptativamente reduz o esforço computacional a cada iteração após analisar localmente e não globalmente o vector de ordenação [43], uma vez que existem páginas em que o índice de ordenação converge mais rapidamente que em outras. No entanto, apesar de na prática este algoritmo adaptativo de PageRank ter provado ser mais eficiente, teoricamente ainda não foi possível demonstrá-lo.

3.3.4 Implementação do algoritmo PageRank

O algoritmo PageRank serve apenas uma parte do sistema de ordenação de páginas feita pelo Google. De facto, o PageRank é incorporado com outros pesos para devolver uma ordenação geral ao sistema [48].
Seguidamente, apresenta-se um modelo básico de implementação e uso do algoritmo Page-

Rank. A sua implementação envolve dois passos principais. Num primeiro passo, é realizado o varrimento completo das páginas para determinar o subconjunto de nodos da rede telemática que contém os termos da consulta. Este subconjunto é designado por universo da consulta e este passo é semelhante ao passo inicial do algoritmo Hits. Num segundo passo, o conjunto é ordenado de acordo com os valores de PageRank associado a cada documento do conjunto.
O cálculo computacional do PageRank é de custo e tempo de execução elevados. Envolve o cálculo do vector estacionário para a matriz estacionária e irredutível, com dimensão na ordem dos biliões, associada à rede telemática. No entanto, e como referenciado por M. Cleuziou e por D. Harman [33, 22], o método descrito garante ser o mais eficiente .
O algoritmo para calcular computacionalmente o vector PageRank π^T, para uma matriz $\overline{\overline{P}} = \alpha\overline{P} + (1-\alpha)\mathbb{1}\mathbb{1}^T\frac{1}{n}$ é simples e é baseado no "Power Method". Após especificar o valor de escolha para α (normalmente $\alpha = 0.85$), e o valor do vector inicial $\pi_0^T = \frac{\mathbb{1}^T}{n}$, onde n representa a dimensão de P, tem-se:

Algoritmo: Cálculo do vector PageRank

Entrada: $n, \alpha, \mathbb{1}$

1: $\pi_0^T = \frac{\mathbb{1}^T}{n}$
2: **while** (o vector nao converge) **do**
3: $$\pi_{k+1}^T = \alpha\pi_k^T\overline{P} + (1-\alpha)\frac{\mathbb{1}\mathbb{1}^T}{n}$$
4: **end while**

Saida: π_{k+1}^T

Na implementação descrita existem certas observações a ter em conta. Primeiro, o método não destrói na totalidade a esparsidade inerente. Em segundo lugar, a multiplicação principal no vector- matriz de $\pi^T\overline{P}$ requer apenas a esparcidade de produtos, e esses produtos são facilmente implementados em paralelo. A utilização de paralelismos é uma forma imperativa de resolver o problema no caso de grandes dimensões. Tem-se verificado que métodos iteractivos mais avançados podem resolver teoricamente e mais rapidamente o problema da convergência deste algoritmo, mas na prática falham devido à complexidade computacional do sua imple-

mentação [33, 22].

Actualmente, tem-se verificado um aumento gradual nos estudos e investigações relativas ao PageRank, em particular, por parte de investigadores de Standford no que diz respeito à implementação do PageRank. Alguns investigadores [1] sugeriram o uso do método de Gaus-Seidel [62] no sentido de simplificar o método do cálculo do PageRank, e verificaram uma rápida convergência do algoritmo, principalmente no início das iteracções. Outro grupo de investigadores desenvolveram diversas modificações ao método para acelerar a convergência. Alterações essas que passaram por aplicar extrapolações quadráticas para acelerar a convergência do vector PageRank. O mesmo grupo de investigadores, desenvolveu o algoritmo BlockRank [57] e um outro algoritmo usando um método adaptativo para monitorização da convergência individual dos elementos do vector PageRank. Outros estudos passam por particionar a matriz P em dois grupos de acordo com a existência ou não de nodos na rede telemática sem apontadores para outras páginas [45, 34, 13].

3.3.5 Exemplo

De seguida apresenta-se um exemplo que demonstra a implementação do algoritmo do PageRank, para solidificação e discussão das ideias gerais do algoritmo e seus resultados. Considere-se o grafo definido na figura 3.6, constituído por um subconjunto de nodos numerados de 1 a 6, onde cada nodo representa uma página de uma rede telemática hipotética.

Pretende-se calcular o vector de ordenação deste grafo e verificar qual das páginas apresenta um valor de π mais elevado, isto é, índice de ordenação mais elevado. A matriz de adjacência associada a este grafo é definida da seguinte forma:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Assim, a matriz de probabilidade P associada ao grafo é dada por

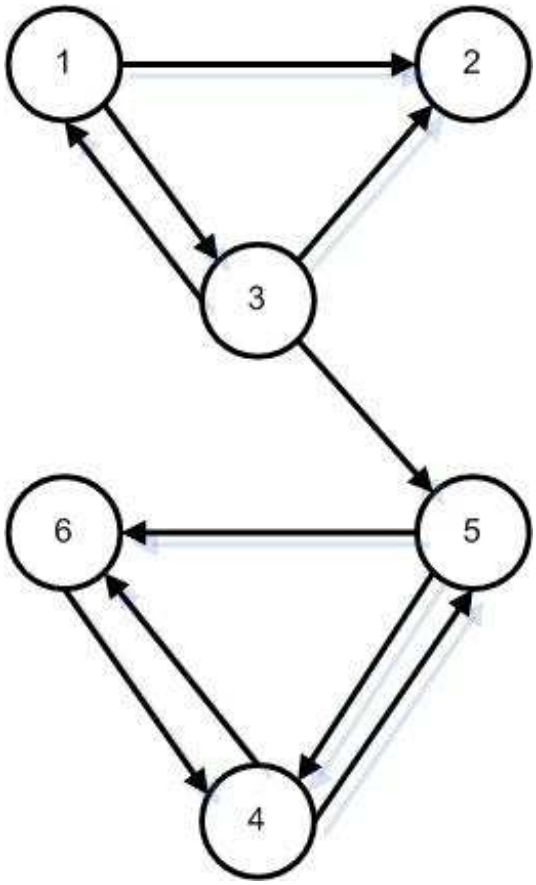


Figura 3.6: Parte da web com 6 páginas 1, 2, 3, 4, 5 e 6

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Esta matriz de probabilidade P, terá de ser modificada até ser uma matriz estocástica e irredutível. Cada uma das linhas que constituem a matriz de transição terão de ser vectores de probabilidade para se poder afirmar que a matriz é estocástica. Todos os estados do grafo deverão comunicar entre si para que a matriz P seja irredutível, como tal não acontece, é necessário torná-la primeiro estocástica e de seguida irredutível. Para que seja estocástica substituem-se todas as linhas nulas da matriz P por $\frac{\mathbb{1}}{6}$ onde $\mathbb{1}$=[1 1 1 1 1 1], obtendo a nova matriz estocástica

$$\overline{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Para ser irredutível, e escolhendo o factor $\alpha = 0.9$ tem-se

$$\overline{\overline{P}} = \begin{pmatrix} \frac{1}{60} & \frac{7}{15} & \frac{7}{15} & \frac{1}{60} & \frac{1}{60} & \frac{1}{60} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{19}{60} & \frac{19}{60} & \frac{1}{60} & \frac{1}{60} & \frac{19}{60} & \frac{1}{60} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{7}{15} & \frac{7}{15} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{7}{15} & \frac{1}{60} & \frac{7}{15} \\ \frac{1}{60} & \frac{1}{60} & \frac{1}{60} & \frac{11}{12} & \frac{1}{60} & \frac{1}{60} \end{pmatrix}.$$

A matriz $\overline{\overline{P}}$ é agora estocástica e irredutível e o seu vector PageRank pode ser calculado usando a equação (3.9) obtendo-se

$$\pi^T = (0.03721,\ 0.05396,\ 0.04151,\ 0.3551,\ 0.206,\ 0.2862)$$

Em concreto, π^T significa e por ordem de relevância decrescente que tem-se

páginas 4 páginas 6 páginas 5 páginas 2 páginas 3 páginas 1.

3.4 Algoritmo SALSA

Desenvolvido por Lempel e Moran [37], o algoritmo Salsa ("*Stochastic Aproach for Link Structure*") combina aspectos do algoritmo PageRank com a ideia desenvolvida pelo algoritmo Hits, em termos de existência de Autoridades e Hubs na estrutura geral da rede telemática. E, tal como acontece nos algoritmos Hits e PageRank, o SALSA cria uma ordenação de Autoridades e de Hubs para as páginas da Internet, baseado no conceito de Cadeias de Markov.
Dado um grafo G, definido a partir do conjunto de páginas da rede telemática resultantes da pesquisa feita pelo utilizador, construa-se um grafo bipartido (ver definição 18), não orientado H, a partir de G.

Definição 18 *Um grafo diz-se* bipartido *se existe uma partição do seu conjunto de nodos em dois outros, $v^{'}$ e $v^{''}$, tal que não existem arcos entre qualquer par de nodos de $v^{'}$ nem entre qualquer par de nodos de $v^{''}$ e pode ser representado por $G = (v^{'}, v^{''}, E)$ onde E é o conjunto de arcos.*

Este novo grafo H é formado por dois conjuntos de nodos, V_a conjunto de potenciais autoridades e V_h conjunto de potenciais Hubs de G.
Depois de serem reordenados, os elementos de V_a e os elementos de V_h passam a formar o conjunto de nodos do grafo H, e como alguns nodos são tanto potenciais autoridades como potenciais hubs, o número total m de nodos de H satisfaz a seguinte condição:

$$m \leq 2n \ , \tag{3.12}$$

onde n é o número total de nodos do grafo G.
Os arcos, não orientados (arestas) de H são definidos através do grafo G como se segue: se G possui um apontador de i para j então coloca-se uma aresta entre o nodo $i \in V_h$ e o nodo $j \in V_a$, e assim sucessivamente até terem sido percorridos todos os caminhos entre os nodos.
Considere-se um nodo de V_a e a partir deste nodo siga-se um caminho aleatório, vai-se ter

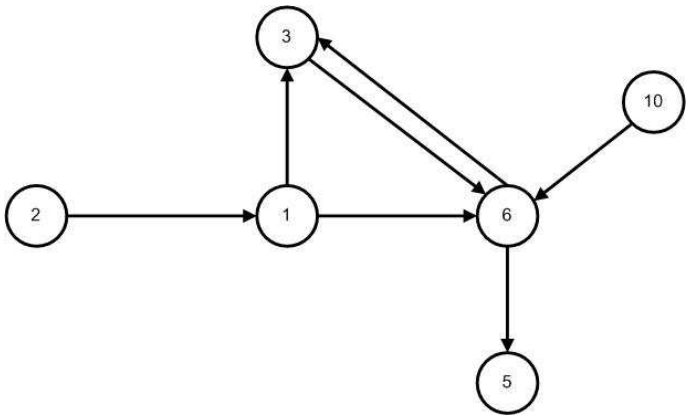


Figura 3.7: Grafo G associado a uma pesquisa na web

ao encontro dum nodo em V_h, e num segundo passo de escolha dum outro caminho aleatório desse nodo de V_h vai-se novamente ao encontro de um novo nodo de V_a e assim sucessivamente. Este par de caminhos aleatórios, um a começar em V_a e outro em V_h pode ser usado para determinar os vectores autoridade e hubs da pesquisa.

Tal como acontece no algoritmo PageRank, cada caminho aleatório é um caminho na Cadeia de Markov, com uma determinada matriz de adjacência associada. Considere-se a matriz A, a matriz de adjacência de G, então W_r será a matriz resultante da divisão de cada entrada de A pela soma das linhas e de forma similar, W_c será a matriz resultante da divisão de cada entrada de A pela soma das suas colunas.

Finalmente obtêm-se os dois vectores de ordenação pelas seguintes fórmulas

$$a_k = W_c^T W_r a(k-1) \quad e \quad h_k = W_r W_c^T a(k-1), \tag{3.13}$$

onde as sequências a_k e h_k dependem da inicialização de a_0 e de h_0, respectivamente.

Neste aspecto, o algoritmo SALSA contrasta com o Hits onde ambas as sequências dependem de h_0, conduzindo a pequenas diferenças na execução dos dois algoritmos. Em particular, o algoritmo SALSA limita a dependência dos vectores a_k e h_k, que nem sempre satisfazem a condição $a_k = W_c^T h_k$. Para melhor se perceber o funcionamento deste algoritmo segue-se um pequeno exemplo.

De maneira similar ao que acontece no algoritmo Hits, inicialmente controí-se o grafo G associado a uma determinada pesquisa na rede telemática. O algoritmo SALSA difere do Hits no próximo passo, antes de definir a matriz de adjacência do grafo G é construído um grafo

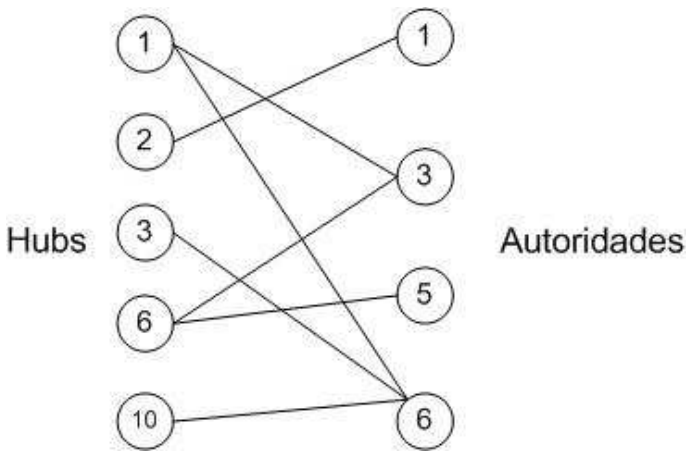


Figura 3.8: Grafo bipartido H de G

bipartido H. O grafo bipartido H é definido por dois conjuntos: V_h, V_a, onde V_h é o conjunto de hubs de G, V_a o conjunto de autoridades de G. Suponha-se o exemplo do grafo bipartido da figura 3.8 construido a partir do grafo da figura 3.7

$$V_h = \{1, 2, 3, 6, 10\} \quad e \quad V_a = \{1, 3, 5, 6\}, \tag{3.14}$$

originando a matriz de adjacência A do grafo G, definida por

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.15}$$

Seja W_r uma matriz idêntica a A mas com as linhas não nulas divididas pela soma do valor de cada linha, e seja W_c uma matriz idêntica a A mas com as colunas não nulas divididas

pela soma do valor de cada, i.e.

$$W_r = \begin{pmatrix} 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{3.16}$$

$$W_c = \begin{pmatrix} 0 & 0 & 1/2 & 0 & 1/3 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \end{pmatrix}. \tag{3.17}$$

Aplicando o algoritmo Hits temos os seguintes vectores de ordenação autoridades e hubs

$$h^T = (.2667,\ \ .2,\ \ .1333,\ \ .2667,\ \ .1333) \quad e \quad a^T = (.25,\ \ .25,\ \ .125,\ \ .375) \tag{3.18}$$

Dado que G é conexo, H e A são Cadeias de Markov irredutíveis, h^T é o vector estacionários de H, que fornece o vector de ordenação de hubs e a^T fornece vector de ordenação de autoridades à pesquisa feita. Caso G não seja conexo, H e A contêm múltiplos componentes irredutíveis logo, os vectores de ordenação global de hubs e autoridades da pesquisa necessitam ambos de serem transformados em vectores estacionários irredutíveis [37].

O algoritmo SALSA, à semelhança do que acontece com o PageRank, resume-se ao cálculo computacional dos vectores próprios associados à matriz de conectividade definida através do grafo G.

3.4.1 Convergência do algoritmo

A convergência do algoritmo é garantida através da transformação da matriz de probabilidades W_c e W_r em matrizes estocásticas, que irão ter apenas um único valor próprio em todo o círculo unitário, pois todos os outros valores próprios têm em módulo valor inferior e muito

próximo da unidade. Gagarante-se assim a convergência do algoritmo para o único vector próprio dominante. No entanto, caso o grafo bipartido H não seja conexo então os vectores de autoridades e hubs não serão únicos.

Teorema 19 *[3] Seja G um grafo orientado com matriz de adjacência A. Os vectores de hubs e autoridades do algoritmo SALSA existem e têm entradas não negativas.*

Prova As matrizes $W_c^T W_r$ e $W_r W_c^T$ são não negativas logo têm sempre um valor próprio $\overline{\lambda}$ positivo, tal que o módulo de todos os outros valores próprios não exceda $\overline{\lambda}$. Um vector próprio associado ao valor próprio $\overline{\lambda}$ pode ser escolhido para que todas as suas entradas sejam não negativas. □

A convergência deste algoritmo é condicionada e pode ser explicada através do seguinte resultado,

Teorema 20 *[3] A sequência a_k proveniente do algoritmo SALSA converge para um vector autoridade, e este vector autoridade é um vector próprio não negativo associado a $W_c^T W_r$. Da mesma forma, h_k converge para um vector hubs e este é um vector próprio não negativo associado a $W_r W_c^T$.*

Prova Os valores próprios de $W_r W_c^T$ são reais e não negativos. Consequentemente a cada iteracção o maior valor próprios em valor absoluto pode repetir-se, mas todos os outros valores próprios são menores em módulo que esse.

Como $W_r W_c^T$ é simétrica, o espaço gerado pelos valores próprios é ortogonal. Deste modo, neste espaço gerado pelos valores próprios pode ser usada a ortonormalização de Gram-Schmidt para escolher os vectores ortogonais tal que pelo menos um deles seja não negativo. Como h^0 é positivo, o produto interno desses vectores ortogonais é positivo e logo a^0 tem uma componente não trivial no espaço gerado pelos valores próprios. Isto implica que o algoritmo converge para um vector próprio associado ao maior valor próprio. Pela construção do algoritmo SALSA, o limite não pode ter entradas negativas. Note-se que embora o algoritmo convirja, pode convergir para qualquer vector não negativo, dependendo da escolha de h^0. O mesmo raciocinio é aplicado para a sequência do vector auroridade. □

3.5 Análise Crítica dos Algoritmos

Todos os algoritmos descritos nas secções anteriores reduzem-se ao cálculo computacional do vector próprio dominante. No entanto, o PageRank e o SALSA calculam o vector próprio recorrendo a conceitos de Cadeias de Markov, enquanto que o algoritmo Hits é formulado em termos do cálculo iteractivo da soma dos pesos de cada nodo.

Uma das grandes vantagens do algoritmo Hits é a apresentação de um vector de ordenação duplo, isto é, apresenta duas listas de ordenação, uma das listas com os documentos do tipo mais autoridades em relação à consulta e a outra com os documentos do tipo mais hub.

O algoritmo Hits não é aplicado à estrutura global da rede telemática, mas apenas a subgrafos desse grafo representativo da estrutura web. Normalmente é aplicado a apenas subgrafos com entre 1000 a 5000 nodos, que derivam directamente da consulta de termos na pesquisa. Este facto implica que o Hits reduz o problema global da pesquisa na rede telemática num problema mais pequeno, pois o custo computacional do cálculo dos vectores próprios associados às matrizes é menor. A dimensão dessas matrizes são muito pequenas comparativamente com o número total de documentos existentes na rede telemática.

Todavia, existem algumas desvantagens do algoritmo Hits. Muitos dos problemas resultam directamente da dependência da consulta de termos. Para cada consulta executada existe necessidade de construir um grafo de vizinhanças, o que origina a necessidade de calcular pelo menos uma matriz de vectores próprios.

O cálculo do vector hub, Y^k e do vector autoridade X^k, com $k \geq 1$, está dependente da escolha feita para o vector hub inicial Y^0. Este factor permite concluir que o algoritmo Hits tem um mau comportamento para certos grafos, pois não consegue retornar um vector de hubs e autoridades único, estão sempre dependentes da escolha da inicialização para Y^0. Por outro lado, o Hits pode devolver uma ordenação de vectores próprios inapropriadamente constituida por entradas nulas. Para certos grafos, o algoritmo Hits poderá originar valores próprios repetidos que condicionam o resultado final. A estes grafos dá-se o nome de grafos mal comportados e o algoritmo nessa situação diz-se mal comportado.

Definição 21 *Um algoritmo de pesquisa diz-se mal comportado, num grafo G qualquer, se o resultado final depende do vector inicial, e/ou se algum nodo do grafo conduz a pesos nulos*

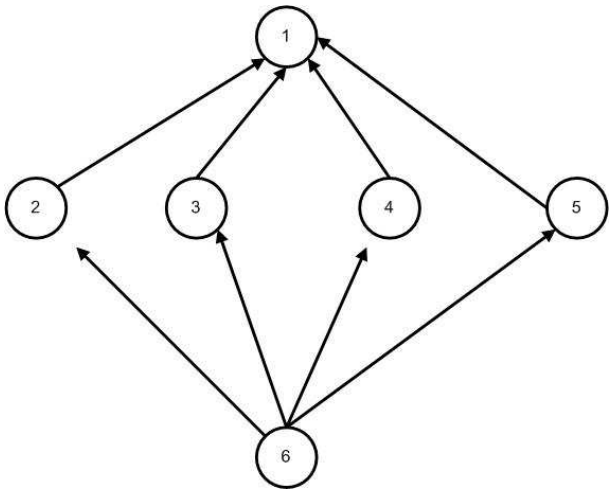


Figura 3.9: Grafo cuja matriz A^TA tem valores próprios repetidos. O algoritmo Hits origina para um determinado input iferentes vectores autoridades, e o SALSA inconsistentes vectores hubs e autoridades.

de autoridades(hub) no vector final.

Pode-se também caracterizar os grafos para os quais o algoritmo Hits é mal comportado através do seguinte resultado

Teorema 22 *[2] Seja A a matriz de adjacência do grafo G. O algoritmo Hits é mal comportado se e só se pelo menos um dos seguintes casos se verifica:*

1. *AA^T contém valores próprios com multiplicidade múltipla;*
2. *O vector próprio associado a AA^T (ou a A^TA) contém entradas nulas para nodos cujos entradas ou saidas são não nulas.*

Considere-se um pequeno exemplo que demonstra claramente este problema de mal comportamento do algoritmo Hits para certos grafos. Primeiro demonstra-se o problema da repetição de valores próprios através da alteração dos valores iniciais.
Através da definição de uma sequência uniforme inicial $Y_0 = [\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}]^T$ para o vector hub, o algoritmo Hits aplicado ao grafo definido na figura 3.9 gera os seguintes vectores hubs e autoridades, $X = [\frac{2}{\sqrt{5}}, \frac{1}{2\sqrt{5}}, \frac{1}{2\sqrt{5}}, \frac{1}{2\sqrt{5}}, \frac{1}{2\sqrt{5}}, 0]^T$ e $Y = [0, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}\ \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}]^T$, respectivamente. Isto significa que o nodo 6 não é melhor hub do que o nodo 2, 3, 4 e 5. De modo

geral, escolhendo vectores iniciais diferentes (positivos) para o algoritmo Hits, o vector de autoridades do grafo da figura será sempre da forma

$$X = [\alpha, \beta, \beta, \beta, \beta, 0]^T,$$

para todo o α e β positivo e tal que $\alpha^2 + 4\beta^2 = 1$. De forma resumida, pode-se afirmar que o algoritmo Hits é um algoritmo instável para certas estruturas da rede telemática. Em particular, o vector de ordenação final depende do vector inicial considerado.

O algoritmo Hits tem uma forte ligação com a pesquisa Bibliométrica. *Din et al.* [45, 34] notaram esta ligação subjacente entre o Hits e dois conceitos comuns de bibliometrias, co-citação e co-referência.

Vários estudos têm sido feitos no âmbito de melhoramento do algoritmo Hits, nomeadamente no que diz respeito à convergência deste algoritmo. O algoritmo Hits Exponencial [16] faz salientar esse melhoramento no que diz respeito à convergência, por meio de pequenas alterações ao algoritmo.

Uma vez que, o algoritmo SALSA foi desenvolvido combinando aspectos dos algoritmos Hits e PageRank tal como acontece com o Hits, também o SALSA devolve ao utilizador resultados duplos (vector autoridade e vector hub), contrariamente ao que se verifica no PageRank que devolve apenas um único vector.

No entanto, um dos problemas do SALSA é a convergência do algoritmo. A dependência do vector inicial e não forçar a irredutibilidade da matriz do grafo, pode originar resultados não únicos. Diz-se assim, pelas mesmas razões que no algoritmo Hits, que o SALSA é um algoritmo mal comportado para certos grafos.

No algoritmo PageRank o grande problema que se coloca é o de determinar um vector de ordenação capaz de mostrar relevância de páginas após a consulta. Muitos trabalhos conduziram a heurísticas para ser aplicadas ao Google no intuito de determinar o vector ordenação de páginas relevantes. De facto, não importa a eficiência do algoritmo se a lista de resultados não fôr relevante para o utilizador. Esta temática gera algumas questões, nomeadamente como medir a relevância de certa página.

A. Langville [35] refere estas desvantagens do PageRank, e.g., o PageRank é independente da consulta realizada e não poder distinguir a popularidade e relevância entre duas páginas. Por outro lado, o uso da popularidade, mais do que a relevância, é a chave principal para o

sucesso do Google ao utilizar o algoritmo PageRank.

Factores como a idade das páginas, números de actualizações feitas e frequência das actualizações às páginas, o uso de apontadores inapropriados, não serão entre outros factores importantes para o cálculo da relevância de uma página? Como medir estes factores quantitativamente? Muitas outras questões podem ser levantadas sobre o PageRank que são alvo de estudo e discussão por parte de investigadores nesta área. Analizando o funcionamento do algoritmo, existe uma grande flexibilidade na personalização ou mesmo intervenção do vector $\mathbb{1}^T$, que o Google é livre de escolher. A escolha de $\mathbb{1}^T$ não afecta o algoritmo em aspectos matemáticos nem computacionais mas poderá afectar de certa forma o indice de ordenação das páginas. Isto poderá ser uma grande vantagem se o Google pretende alterar as posições de PageRank das páginas na rede telemática, ou por interesses económicos ou talvez devido a suspeitas de apontadores inaproprados em determinadas páginas.

Mas existem alguns aspectos a salientar na implementação do algoritmo. Primeiro que tudo, verifica-se que este algoritmo não faz uma ordenação global das páginas da rede telemática mas sim de cada página individualmente. O PageRank de uma certa página A é definido recursivamente através de todas as páginas que têm apontadores a apontar para a página A. Este algoritmo não tem em conta qualquer tipo de palavra-chave, que poderia ser adicionada no cabeçalho de cada página, contribuindo assim para alguma ambiguidade entre as pesquisas englobando termos na pesquisa que não os pretendidos. Atribuindo apontadores descontextualizados do objectivo da página poderá alterar o índice de ordenação de uma página, modificando os resultados da pesquisa, induzindo a resultados tendenciosos e irrelevantes.

Um exemplo recente disso é a pesquisa por “failure“ ou “miserable failure“ que retorna como primeiro site a biografia oficial da Casa Branca para George W. Bush, presidente dos Estados Unidos, e em seguida a página de Michael Moore, inimigo declarado do presidente dos EUA. Este processo ficou conhecido por “Googlebombing“.

De facto, a existência de replicação de páginas em bases de dados onde o algoritmo não consegue distinguir a sua identidade poderá modificar a ordenação final da pesquisa. Veja-se o seguinte exemplo que poderá explicar por si só algumas das patologias deste algoritmo.

Dado o grafo conexo G na figura 3.10, representativo de uma estrutura da rede telemática cuja matriz de adjacência é definida por

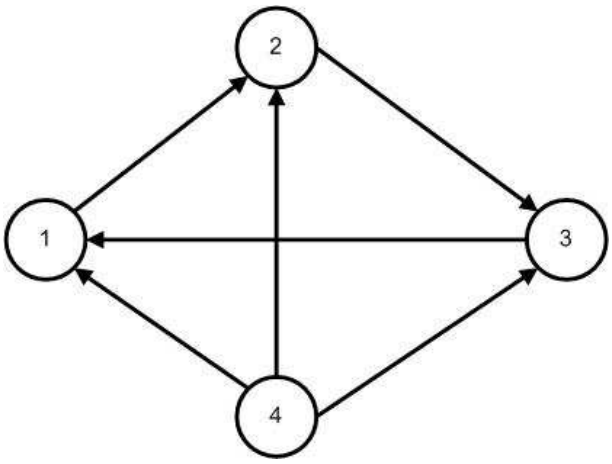


Figura 3.10: Grafo associado a uma estrutura web

$$\mathrm{A}=\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix},$$

Obtem-se por aplicação do algoritmo do PageRank o vector de ordenação definido por

$$\pi^T = \begin{pmatrix} 0.3102 & 0.3102 & 0.3102 & 0.0694 \end{pmatrix}$$

Note-se que a página com menos relevância é a página identificada pelo nodo 4, pois é a que apresenta menor valor de índice de ordenação. Agora adicionem-se duas novas páginas, 5 e 6, onde a página 6 tem um apontador para a página 5 e a página 5 é ligada à página com maior índice de ordenação, por exemplo a página 1. Introduza-se também um apontador da página com menor índice de ordenação, neste caso a 4, para a página 5, resultando o grafo definido na figura 3.11 cuja matriz de adjacência é dada por

$$\mathrm{A}=\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

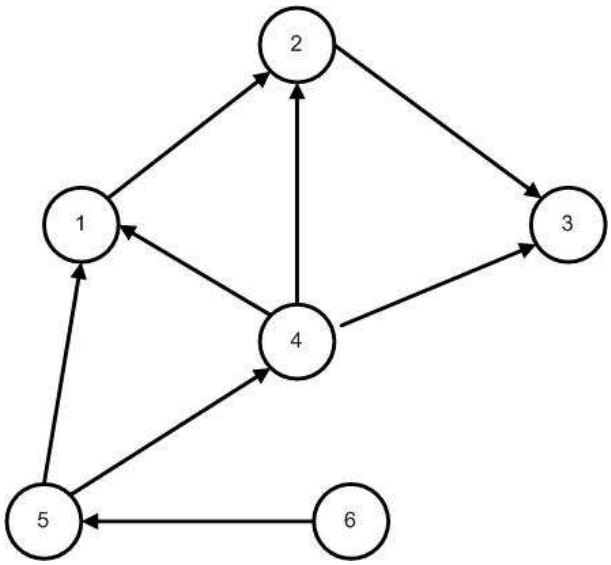


Figura 3.11: Grafo após a adição de novos vértices

a qual continua a representar um grafo conexo. Aplicando novamente o algoritmo PageRank resulta que o vector de ordenação é

$$\pi^T = \begin{pmatrix} 0.2990 & 0.2706 & 0.2499 & 0.0463 & 0.0880 & 0.0463 \end{pmatrix}$$

onde se observa que a página 4 perde importância relativamente às páginas que foram adicionadas nomeadamente à 6 que está ligada somente à 5, passando a estar em último lugar, com menor índice de ordenação.

Poder-se-á perguntar, como é que a introdução de um novo apontador na estrutura de uma página i para uma página j poderá afectar a ordenação dessas duas páginas. A resposta a esta questão poderá ser tratada utilizando ferramentas da teoria de Cadeias de Markov.

Lema 23 *[5] Para todo $i, j \in \{1, ..., n\}$ com $i \neq j$, se à página i for adicionado um novo apontador para a página j então o PageRank da página j aumenta.*

Prova: Para uma cadeia de Markov $\{X_k, k \geq 0\}$ com espaço de estados $\{1, 2, ..., n\}$, matriz de transição $\overline{P}$ e para qualquer $m > 0$, definam-se as seguintes probabilidades:

$$f_{ij}^{(m)} = P(X_m = i, X_k \neq j, X_k \neq j, 1 \leq k < m | X_0 = j),$$
$$f_{ji}^{(m)} = P(X_m = i, X_k \neq i, X_k \neq j \ 1 \leq k < m | X_0 = j),$$
$$g_{ij}^{(m)} = P(X_m = i, X_k \neq j \ 1 \leq k < m | X_0 = i).$$

Observe-se que, por exemplo, o valor $f_{ij}^{(m)}$ é a probabilidade de começando na página j escolher um caminho aleatório e alcançar a página i, a primeira vez, após m passos, sem voltar a visitar a página j. Da mesma forma, $g_{ij}^{(m)}$ é a probabilidade, de começando na página i escolher um caminho aleatório e alcançar a página j depois de m passos.
O valor μ_j corresponde à matriz original $\overline{P}$, sem o novo apontador de i para j, sendo dado pela a soma dos três termos:

$$\mu_j = \sum_{m=1}^{\infty} m f_{jj}^{(m)} + \sum_{m=1}^{\infty} m f_{ji}^{(m)} + \sum_{m=1}^{\infty} m g_{ij}^{(m)}$$

O primeiro termo do segundo membro da igualdade corresponde à contribuição dos caminhos que começam em j e regressam a j sem passar por i. O segundo termo corresponde à soma de todos os caminhos que começam em j e acabam em i no primeiro passo. Finalmente, o terceiro termo corresponde a todos os caminhos que começam em i e acabam em j.
Agora, suponha-se que um novo apontador é adicionado de i para j. Com a nova matriz, o primeiro termo permanece inalterado pois representam todos os caminhos que não passam em i, o mesmo é válido para o segundo termo. No caso do terceiro termo, verifica-se de imediato que vai diminuir o seu valor pois, apesar do comprimento dos caminhos que não estão a usar o novo apontador permanecer inalterado, a probabilidade desses caminhos reduz-se devido ao peso dado ao novo apontador, mesmo que seja pouco. Logo, o novo apontador vai fazer reduzir o valor de μ_j. Então, a ordenação obtida é certamente alterada, mais para a página j.□

As pesquisas e variantes do algoritmo PageRank são diversas, no sentido de melhorar o algoritmo original criado pelos fundadores do Google. Na tentativa de perceber a essência do PageRank foram sugeridos por O. Nogueira [16] um conjunto de axiomas para explicar a perspectiva da escolha do utilizador em lidar com a relevância das páginas através do PageRank. Cada página pertencente ao grafo da rede telemática é visto como um agente, onde se esse agente prefere outros agentes (i.e., outras páginas) então contém apontadores para eles. É apresentado por L. Yang [29] o "Weighted PageRank" algoritmo (WPR), uma extensão do algoritmo PageRank. O WPR têm em conta a importância tanto dos apontadores que saem como os que entram em determinada página e distribui os resultados da ordenação baseado na popularidade das páginas. Este algoritmo apresenta resultados computacionais melhores que o PageRank para uma determinada consulta executada, mas é computacionalmente mais

exigente.

Outras variantes existem que, de alguma forma tentam melhorar o PageRank, como o caso do "Predictive Ranking" (PreR) e o "BackRank" descritas por M. Bouklit [39].

A tentativa de implementação dos algoritmos Hits, SALSA e PageRank sobre base de dados relacionais é explicada por M. G. Diligenti [40], i. e., quando as páginas são substituidas por dados num servidor SQL.

3.6 Efeito de Apontadores Inapropriados no Algoritmo PageRank

Nesta secção ir-se-á estudar outro tipo de incentivo para os utilizadores da rede telemática, de classificar apenas páginas relevantes relacionadas e agrupadas coerentemente. Nomeadamente, vai-se identificar e analisar as consequências da existência de apontadores inapropriados. Define-se por apontadores inapropriados todos os apontadores que apontam de um grupo de páginas da rede telemática para outro, no entanto, os dois grupos têm conteúdos consideravelmente distintos. Esperar-se que o número de tais apontadores seja mais pequeno que o número de apontadores que ligam documentos pertencentes ao mesmo grupo e que muitas das vezes esses apontadores inapropriados ligam um ponto do grupo a outro mas apenas num dos sentidos. Para modelar tal situação, considerem-se dois grupos, grupo 1 e grupo 2, com alguns apontadores do grupo 2 para o grupo 1 mas nenhum apontador do grupo 1 para o grupo 2. Sem a existência de apontadores inapropriados a matriz de transição de estados pode ser definida por

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}, \tag{3.19}$$

onde P_1 e P_2 são as matrizes de transição do grupo 1 e grupo 2, respectivamente.

No entanto, com a adição de apontadores inapropriados a matriz de transição passa a ter seguinte forma

$$\overline{P} = \begin{pmatrix} P_1 & 0 \\ R & P_2' \end{pmatrix}, \tag{3.20}$$

onde R é a matriz que representa as ligações do grupo 2 para o grupo 1. Note-se que P_1 e P_2 são matrizes estocásticas e $P_2^{'}$ é uma matriz sub-estocástica. De seguida apresenta-se o estudo de dois casos: um caso geral, quando qualquer página do grupo 2 contém apontadores inapropriados; e uma caso particular, quando apenas uma página do grupo 2 têm apontadores inapropriados.

3.6.1 Caso geral: diversas páginas com apontadores inapropriados

Teorema 24 *[5] Seja $[\pi_1\pi_2]$ o vector PageRank da matriz $\overline{P} = \alpha P + (1-\alpha)\frac{\mathbb{1}}{n}E$. E seja $[\pi_1^{'}\pi_2^{'}]$ o vector PageRank da matriz da rede telemática com apontadores inapropriados definida em (3.20). Então $\pi_1^{'}$ e $\pi_2^{'}$ podem ser definidos da seguinte forma*

$$\pi_2^{'} = \frac{1-\alpha}{n} 1^T [I - \alpha P_2^{'}]^{-1}, \tag{3.21}$$

$$\pi_1^{'} = \pi_1 + \alpha \pi_2^{'} R [I - \alpha P_1]^{-1}. \tag{3.22}$$

Prova Através da ideia subjacente de decomposição de matrizes, pode-se provar as afirmações directamente através da fórmula

$$[\pi_1^{'}\pi_2^{'}][I - \alpha P^{'}\ ] = \tfrac{1-\alpha}{n}\mathbb{1}^T$$

escrita na forma de bloco. □

3.6.2 Caso particular: uma página com apontadores inapropriados

Pode-se fazer um estudo mais profundo da adição de apontadores inapropriados através da análise do caso em que apenas uma das páginas do grupo 2 contém apontadores inapropriados. Suponha-se que uma página $i \in \{1, ..., n_2\}$, com n_2 o número total de páginas do grupo 2, inicialmente tem k_2 apontadores para as páginas do grupo 2 e de seguida adicione-se

k_1 apontadores ao grupo 1. Mais ainda, admita-se que as outras páginas do grupo 2 têm apontadores para apenas uma outra página. Denote-se $k = k_1 + k_2$. Seja U uma matriz $m \times n_2$, onde m é o número de páginas do grupo 2 com apontadores para o grupo 1, com m linhas não nulas de $P_2 - P_2^{'}$. Neste caso, U é igual a um vector linha u com número de entradas não nulas igual a $\frac{1}{k_2} - \frac{1}{k} = \frac{k_1}{kk_2}$. Observe-se que u não é mais do que a i-ésima linha de P_2 multiplicada por $\frac{k_1}{k}$. O vector v^T é agora igual ao vector coluna e_i. Neste caso, resulta a seguinte formula,

$$\pi_2^{'} = \pi_2 - \frac{\alpha\pi_2^i u[I - P_2]^{-1}}{+\alpha u[I - P_2]^{-1}e_i}, \tag{3.23}$$

onde π_2^i é a i-ésima coordenada de π_2.

Desta forma, a partir da fórmula anterior pode inferir-se a probabilidade do grupo 2 perder valor de PageRank quando uma página ligada a essa contém apontadores inapropriados. Multiplicando ambos os membros da formula pelo vector coluna $\mathbb{1}$, com todas as entradas iguais a 1, e tendo em conta que

$$[I - \alpha P_2]^{-1}\mathbb{1} = \sum(\alpha P_2)^l \mathbb{1} = [1/(1-\alpha)]\mathbb{1}$$

obtém-se

$$(\pi_2\mathbb{1} - \pi_2^{'}) = \frac{\alpha k_1 \pi_2^i}{(1-\alpha)k(1 + \alpha u[I - \alpha P_2]^{-1}e_i}. \tag{3.24}$$

Este conceito de grupos foi também utilizado por G. M. Scarselli [38], onde a probabilidade é interpretada como uma energia, e expressa a perda de energia em termos do novo PageRank $\pi_2^{'}$.

Um caso interessante é quando uma determinada página i tem apontadores para todas as páginas do grupo 2, incluindo para ela própria. Neste caso, tem-se

$$u = \frac{k_1}{kk_2}\mathbb{1}^T$$

e a equação (3.23) passa a representar-se na forma

$$\pi_2^{'} = \pi_2(1 - \frac{\alpha k_1 \pi_2^i}{kk_2 + \alpha k_1 \pi_2^i}). \tag{3.25}$$

Observe-se que, neste caso, o valor índice de ordenação perdido pela página i do grupo 2 é proporcional a π_2^i.

De acordo com a matriz de transição $\overline{P} = \alpha P + (1-\alpha)\frac{1}{n}\mathbb{1}$, a cada iteracção, a hiperligação

de transição da matriz P ocorre com probabilidade α e a transição aleatório pela matriz $\frac{1}{n}E$ ocorre com probabilidade $(1-\alpha)$. Agora considere-se o caminho simples de transição de hiperligações no grupo 2 até que a primeira transição ocorra. Denotando por z_{ij} o número de visitas à página j tal que o caminho até ele tenha começado em i, quando definidas os números de transições através da matriz αP_2. Note-se que z_{ij} é igual à entrada (i,j) da matriz $[I-\alpha P_2]^{-1}$, e tal que

$$z_{ij} = e_i^T[I-\alpha P_2]^{-1}e_j \text{ com } i,j \in \{1,....,n_2\}.$$

Assim sendo, pode-se aplicar a seguinte interpretação

$$\pi_2^j - \pi_2^{'j} = \frac{\pi_2^i VP_{ij}}{1+VP_{ij}} \tag{3.26}$$

para $j \in \{1,....,n_2\}$, $VP_{ij} = E(\sharp$ visitas perdidas de i para j) onde VP_{ij} representa o número esperado de visitas perdidas de i para j. No seguinte teorema é apresentada uma forma de determinar o valor de indice de ordenação de PageRank perdido pela página i.

Teorema 25 *[5] Se uma página $i = 1,2,...,n_2$ do grupo 2 tem k_1 ligações para o grupo 1 e k_2 ligações para o grupo 2, e todas as páginas do grupo 2 tem ligações apenas entre elas, então*

$$\pi_2^i - \pi_2^{'i} = \frac{\pi_2^i z_{ii} + k_2 - k}{k_1 z_{ii}^2 + k_2}. \tag{3.27}$$

Prova Para provar o resultado descrito, basta apenas determinar o número esperado de visitas perdidas de i para i. É trivial verificar que,

$$VP_{ii} = VP_{ii}(\alpha P_2) - VP_{ii}(\alpha P_2^{'}), \tag{3.28}$$

onde $VP_{ii}(\alpha P_2) = E(\sharp$ " visitas perdidas de i para i com αP_2 ")
e
$VP_{ii}(\alpha\ P_2^{'}) = E(\sharp$ " visitas perdidas de i para i com $\alpha P_2^{'}$ ").
Agora, seja q_{ii} a probabilidade de fazer uma transição aleatória de i para i com αP_2). Então, pelas propriedades das Cadeias de Markov

$$[\text{ " número total visitas de i para i com } \alpha P_2 \text{ "}] + 1 \tag{3.29}$$

tem uma distribuição geométrica com parâmetro q_{ii}. Pela mesma razão

$$[\text{ " número total visitas de i para i com } \alpha P_2^{'} \text{ "}] + 1 \tag{3.30}$$

segue uma distribuição geométrica com parâmetro q'_{ii}. Então tem-se

$$q_{ii} = z_{ii}^{-1}. \tag{3.31}$$

Por outro lado,

$$1 - q_{ii} = \alpha P,$$

com $P = P($ " alcançar i a partir de i com α P_2"| " a 1ª transição seguiu com αP_2") Denotando esta última probabilidade condicionada por γ_{ii}, tem-se que

$$\gamma_{ii} = (1 - z_{ii}^{-1})/\alpha.$$

Mais ainda, pode-se observar que

$$1 - q'_{ii} = \alpha(1 - \frac{k_1}{k})\overline{P} = \alpha(1 - \frac{k_1}{k})\gamma_{ii} = \frac{k_2}{k}(1 - z_{ii}^{-1}) \tag{3.32}$$

onde $\overline{P} = P($" alcançar i apartir de i com α P'_2"| " a 1ª transição seguiu com $\alpha P'_2$") e onde a segunda equação só é válida desde que P'_2 defira de P_2 apenas na sua i-ésima linha. Assim sendo resulta que,

$$\alpha u[I - \alpha P_2]^{-1} e_i = \frac{1}{q_{ii}} - \frac{1}{q'_{ii}} = z_{ii}(1 - \frac{k}{k_1 z_{ii} + k_2}) \tag{3.33}$$

Substituindo (3.33) em (3.23) ou em (3.26) chega-se então ao resultado pretendido. □

Capítulo 4

Aplicação do PageRank a um Catálogo Científico

4.1 Mecanismo de Pesquisa

A investigação científica depende maioritariamente de pesquisa e consulta de documentos e material bibliográfico de longos períodos de tempo. No entanto, com o aumento permanente e sem regras da literatura científica torna-se difícil para os autores, revisores e utilizadores identificarem e utilizarem toda a informação relevante no seu trabalho. Verifica-se ainda hoje que os documentos publicados em jornais científicos são maioritariamente caracterizados pelo seu título, autores, palavras-chave, dados do jornal e ano de publicação no sistema, no entanto estes metadados continuam a ser insuficientes para o sistema devolver dados relevantes à consulta de um utilizador. Os metadados são muitas das vezes insuficientes como recurso nos processos de pesquisa em catálogos científicos, contudo continuam a ter um papel fundamental na interligação, identificação, gestão e preservação da pesquisa nos sistemas usuais. São fundamentais para um vasto grupo de pessoas: investigadores quando necessitam de encontrar informação relevante ao seu trabalho; editoras que pretendem divulgar o seu produto; centros e locais de digitalização de documentos para organizar o seu trabalho; catálogos cientificos ou arquivos de informação [11].

Encontrar a informação certa para determinada pesquisa pode tornar-se difícil, assim a solução poderá passar por encontrar mecanismos ou fazer melhoramento aos mesmos, resolvendo assim o problema da desproporcionada quantidade de informação em catálogos científicos. Em geral, tem-se observado que os artigos científicos mais citados são os de mais fácil acesso [20], no entanto nada garante que sejam os mais relevantes. Pode-se garantir, pelo menos, que serão os mais populares numa determinada área de investigação.

Um exemplo de um mecanismo tradicional de pesquisa é a pesquisa por autores de artigos científicos, no entanto, claramente inadequado quando a lista de autores a procurar é grande, ou simplesmente desconhecemos os autores relevantes. Por isso, facilitar e refinar esse mecanismo poderá ser de grande importância para uma apresentação de resultados mais próxima possível do que o utilizador pretende. Outro exemplo de um mecanismo de pesquisa é considerar a pesquisa por palavras-chave a serem consideradas pelo sistema quando devolve ao utilizador um conjunto de documentos relevantes.

O objectivo principal deste capitulo é aplicar um algoritmo de pesquisa baseado em apontadores, nomeadamente o PageRank, a um catálogo científico. Este algoritmo poderá ser interpretado como um novo método de ordenação de artigos mais "relevantes", i.e., mais populares no catálogo.

Considere-se um catálogo científico, C, constituído por um conjunto de documentos científicos. Cada documento é caracterizado por atributos, i. e., os metadados.

Um documento poderá ter um ou mais autores e um mesmo autor poderá ser citado em documentos distintos. Um documento só poderá citar um ou mais documentos que tenham sido já publicados, i.e., só poderá citar documentos anteriores a ele, ímplicando a não existência de ciclos no grafo associado. Quanto mais documentos citarem um determinado documento mais popular esse documento é, por outro lado, quanto mais documentos um determinado autor possuir mais popular é este autor.

Poderão surgir dois tipos de abordagem a este problema de pesquisa de documentos: numa das abordagens, a pesquisa é feita directamente aos documentos e aplicado directamente o algoritmo PageRank aos documentos do catálogo; na outra abordagem, será considerado que cada documento é constituido por um ou mais autores e que esses autores citam outros autores, podendo então obter-se um vector de ordenação dos autores mais populares, utilizando o algoritmo PageRank.

4.2 Pesquisa de Documentos por Aplicação do PageRank

Seja C um catálogo científico e $D = \{d_i : i = 1, ..., N\}$ o conjunto de todos os documentos de C. Para cada documento d_i poderão existir um ou mais autores a citá-lo, logo ter-se-à um ou mais apontadores de outros documentos para este documento. Deste modo, pode-se contruir um grafo orientado $G = (V, E)$, onde V é o conjunto de todos os documentos de C e E o conjunto de todas as citações entre os documentos. Como referido anteriormente, só existem apontadores de um artigo para artigos anteriores a este e também não são consideradas citações a trabalhos que tenham o mesmo autor. Deste modo, o grafo G não tem laços nem ciclos.

Cada documento do conjunto V pode ser visto como uma autoridade ou como um hub. Uma autoridade é um documento que tem outros documentos a apontarem para ele, i.e., que tem outros documentos a citá-lo, em contrapartida um hub é um documento que aponta para outros documentos, i.é., um documento que cita outros documentos, ver figura 4.1.

A partir desta informação poderá ser construido o grafo associado aos documentos do catálogo e os seus apontadores devem ser assinalados. O algoritmo de PageRank é directamente aplicado a este caso e como resultado obtém-se o vector PageRank π dos documentos mais populares do catálogo para a consulta realizada.

4.3 Pesquisa de Autores por Aplicação do PageRank

O propósito do processo de pesquisa em catálogos científicos é suportado pela pesquisa de documentos mais ou menos populares para determinada consulta. O mecanismo de pesquisa baseado em autores de documentos pode ser útil e vantajoso quando se pretende saber que autores têm as mesmas áreas de investigação de interesse, ou de modo geral, quando se pretende saber quais os autores mais populares no sistema. Quando se define que um autor é mais popular que outro pode-se estar a afirmar que esse autor publicou mais artigos na área, ou o mais citado por outros autores da mesma área de investigação. Aplicar o algoritmo

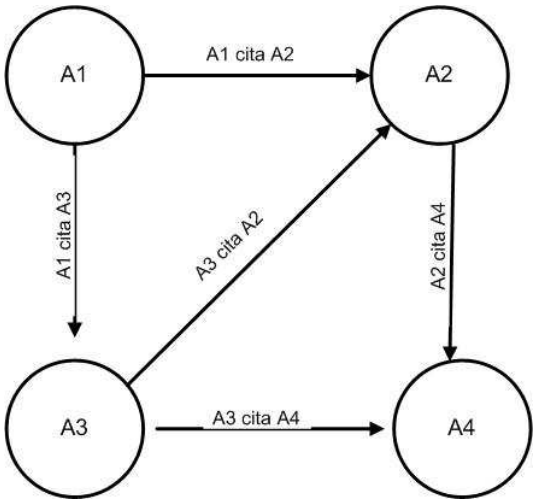


Figura 4.1: Grafo onde cada nodo é um artigo e cada arco uma citação

PageRank, baseado em estruturas de apontadores, à pesquisa de autores é o propósito desta secção.

Seja D o universo de documentos do catálogo científico C e A o universo de autores do catálogo. Com a informação da citação de um documento para outros documentos, conforme observado na secção anterior, pode-se construir o grafo de citações $G = (D; E)$, onde D é o conjunto de documentos de C e E os arcos entres os documentos, $E = \{(i, j) : \ documento \ i \ cita \ o \ documento \ j\}$.

Cada documento é constituído por um ou mais autores, esse documento cita um ou mais documentos que por sua vez têm um ou mais autores, veja-se a figura 4.2. Com base nesta informação, pode-se construir um grafo H, que representa todas as ligações entre os autores dos documentos, tendo por base de construção as citações. Um documento d_i tem um conjunto de autores $\{a_1, a_2, ..., a_k\}$ com $k \in \mathbb{N}$, d_i cita o documento d_j com $i \neq j$, pois não se consideram nem documentos que não citam qualquer outro documento nem que tenham citações a eles próprios. O documento d_j tem um conjunto de autores $\{b_1, b_2, ..., b_n\}$ com $n \in \mathbb{N}$. Dado que d_i cita d_j então pode-se colocar um apontador de cada um dos autores de d_i para todos os autores de d_j e assim sucessivamente com todos os outros documentos. O grafo que se obtém é o grafo H orientado que tem todas as ligações entre todos os autores que citam e são citados por outros autores. Os nodos de H são os autores e os arcos as citações. Este grafo representa uma rede telemática, onde os apontadores são as citações entre autores. Construído o grafo

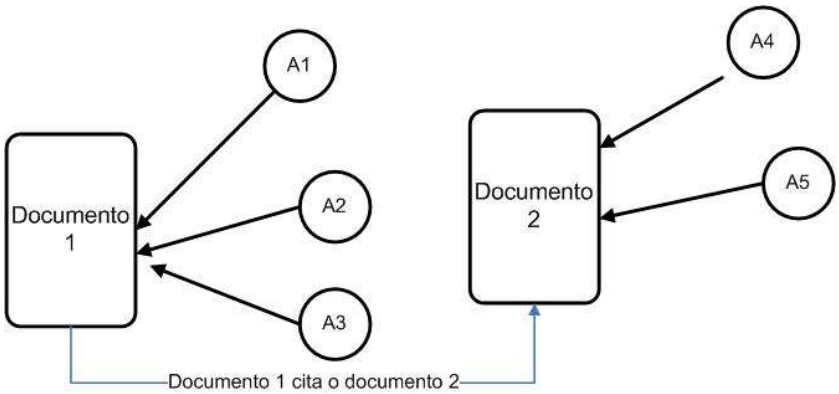


Figura 4.2: Esquema onde se observam as ligações entre documentos e os autores que cada um deles têm

está-se em condições de aplicar o algoritmo PageRank a este e assim, obter uma lista ordenada dos autores mais populares do catálogo científica para determinada consulta. De seguida, é exposto um exemplo que mostra na prática a aplicação deste algoritmo na pesquisa de autores em catálogos científicos.

Considere-se o grafo G, representado na figura 4.3, construído a partir dos documentos da consulta do catálogo científico. Neste grafo estão descritas todas as ligações possiveis entre os 5 documentos construídas a partir das citações. Para cada um dos 5 documentos também se sabe quais os autores que são citados e dai facilmente se constroí o grafo H que representa todas as ligações que existem entre os autores dos 5 documentos como mostra a figura 4.4. Observa-se que cada documento i tem os autores A_{ij}, onde $i \in \{1, ..., 5\}$, $j \in \{1, 2\}$ e j depende de i. Por exemplo, o documento d_1 tem dois autores, o A_{11} e o A_{12}, o documento d_2 tem apenas um autor, A_{21}. A partir deste grafo constroi-se a matriz de adjacência associada a H, onde A é dada por:

$$\mathrm{A}=\begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

A matriz de probabilidade P associada ao grafo é definida da seguinte forma

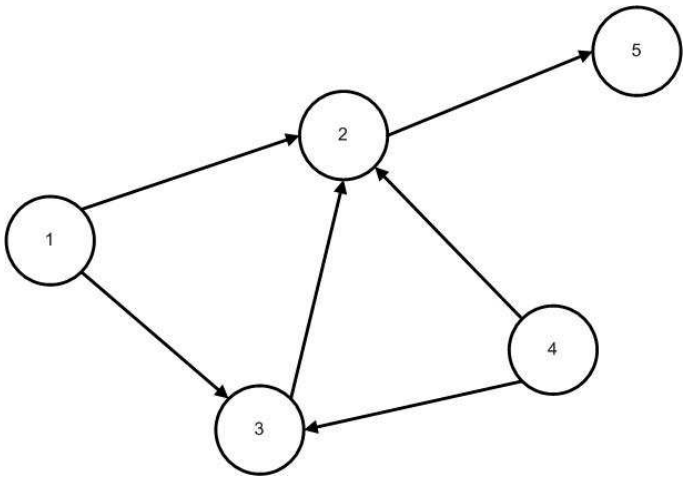


Figura 4.3: Grafo com as citações de 5 documentos

$$\mathrm{P}=\begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Esta matriz de probabilidade P tem de ser estocástica e irredutível para poder aplicar o "Power Method" associado ao algoritmo PageRank. Todos os estados do grafo têm de comunicar entre si para que a matriz ser irredutível e para cada uma das linhas da matriz a soma dos seus elementos tem de ser igual a um para a matriz ser estocástica. Como esta matriz não é estocástica nem irredutível tem de se forçar a irredutível da matriz e fazer com que seja estocástica. Para que seja estocástica basta substituir as linhas nulas de P por $\frac{\mathbb{1}}{6}$. A nova matriz estocástica é definida por

$$\overline{P}=\begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

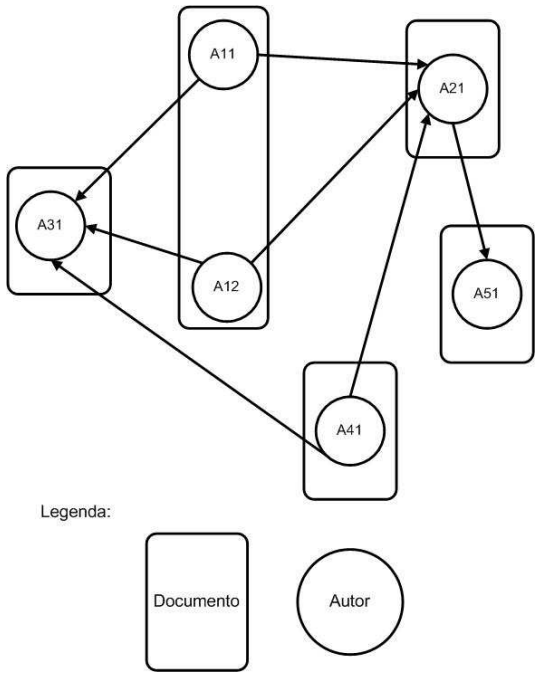


Figura 4.4: Grafo com os autores de cada um dos documentos e ligações entre eles

Basta agora tornar $\overline{P}$ irredutível. A nova matriz $\overline{\overline{P}} = \alpha\overline{P} + (1-\alpha)\frac{\mathbb{1}\mathbb{1}^T}{6}$ onde $\alpha = 0.85$ é a matriz irredutível que se pretendia obter e é dada por

$$\overline{\overline{P}} = \begin{pmatrix} 0.1500 & 0.1500 & 0.5750 & 0.5750 & 0.1500 & 0.1500 \\ 0.1500 & 0.1500 & 0.5750 & 0.5750 & 0.1500 & 0.1500 \\ 0.1500 & 0.1500 & 0.1500 & 0.1500 & 0.1500 & 1.0000 \\ 0.2917 & 0.2917 & 0.2917 & 0.2917 & 0.2917 & 0.2917 \\ 0.1500 & 0.1500 & 0.5750 & 0.5750 & 0.1500 & 0.1500 \\ 0.1500 & 0.1500 & 1.0000 & 0.1500 & 0.1500 & 0.1500 \end{pmatrix}.$$

A partir deste ponto têm-se todas as condições necessárias para aplicar o algoritmo de PageRank, obtendo-se a ordenação de popularidade dos autores. O vector de ordenação, vector PageRank π, obtido é

$$\pi = [0.0617, 0.0617, 0.3625, 0.1287, 0.0617, 0.3236].$$

Facilmente se verifica que o autor mais popular deste grafo é o autor A_{21}, autor que pertence ao documento d_2. Todas as características e propriedades do algoritmo PageRank em estru-

turas de apontadores é mantida neste modelo, nomeadamente a convergência e o número de iteracções necessárias.

Capítulo 5

Um Novo Modelo de Pesquisa para Catálogos Científicos

5.1 Pesquisa Personalizada

A Internet surge da possibilidade de criar um sistema, ambiente, não controlável nem controlado de informação. No entanto, com os progressos cientificos e tecnológicos, a Internet, não passa maioritariamente de um arquivo de informação sem a utilidade desejada. Por exemplo, um utilizador assíduo da rede telemática poderá ler, das cerca de 15 milhões de páginas existentes, no máximo 100, por cada pesquisa que realiza. No entanto, existe a tentativa permanente por parte das instituições estatais e internacionais, e também de investigadores científicos nas universidades, de catalogar e desenhar novas topologias de acesso e percepção da Internet como uma rede conexa de informação organizada e ligada por hiper ligações, que vai ao encontro dos requisitos básicos de cada utilizador. Veja-se o caso do W3C e da Web 2.0.

Os obstáculos que se encontram, na tentativa permanente de gerir a informação inserida na rede telemática são muitos e variados. A maioria destes obstáculos surge da necessidade de fazer um mapeamento da informação inserida da rede telemática. Por exemplo, não saber como se encontra estruturada a informação quando se pretende fazer a sua gestão na rede é

um dos principais obstáculos existentes. A presença de diferentes maneiras de catalogação usadas pelos catálogos científicos é outro dos obstáculos com que se pode deparar.

Até há pouco tempo, a informação na rede telemática era apenas textual, com evolução do conhecimento a esta tem sido acrescentado o som, imagens e filmes. Por isso, a catalogação de documentos passa hoje em dia por incluir também o processamento de informação digital, electrónica e com linguagens específicas. Todos estes factos interferem de uma maneira geral na forma como um catálogo clarifica a informação que posteriormente disponibiliza para o utilizador.

A diversidade da construção destes sistemas implica, resultados diferentes em termos de pesquisa, e logo, em termos de avaliação de relações de dependência contextual entre os documentos.

Existem vários estudos realizados sobre catálogos científicos, a sua eficácia, características e modos de funcionamento, mas de uma maneira geral todos estes sistemas podem ser definidos como sistemas de recolha de informação (IR) que contêm um processo de recolha de documentos e um de perguntas, contêm funções de pesquisa e comparação da dados, uma pergunta e a possibilidade de visualizar a informação de acordo com uma ordenação previamente definida. As avaliações a estes sistemas são feitas geralmente em termos de pesquisa, quantidade de documentos indexados, velocidade da recolha, e hipóteses de cingir as respostas ao tema da pergunta.

Existe uma grande comunidade de catálogos científicos na rede telemática para organização da informação, mas pretende-se definir e caracterizar nas próximas secções um catálogo científico especializado em documentos na área da Matemática. MathSciNet, Zentralblatt, Scopus, ou mesmo, o Scholar Google são os principais sistemas de pesquisa de informação existentes na rede telemática para documentos nessa área, entre outros. O sistema de pesquisa da Zentralblatt [68] contém desde 1868 mais de 2.0 milhões de entradas, o sistema MathSciNet [42] tem a mesma ordem de grandeza e são adicionados todos os anos itens, dá cobertura a cerca de 1799 jornais da comunidade científica e possui ligações da rede telemática para cerca de 717220 artigos científicos de matemática. Scopus [59] é o sistema de pesquisa e armazenamento de informação com maior abstracção e qualidade nos dias de hoje para as várias ciências. Diariamente actualiza os seus dados e oferece informação de cerca de 12.850 jornais académicos e 28 milhões de artigos científicos, nas mais diversas áreas.

Conseguir construir um sistema de armazenamento, pesquisa e visualização de informação científica na área da Matemática, que tenha como principal factor de apresentação a relevância que o utilizador dá aos documentos e o perfil definido por este, foi o principal objectivo deste trabalho.

Acontece frequentemente que quando um utilizador interessado em artigos de matemática faz uma pesquisa na rede telemática seja surpreendido por resultados que não vão de encontro ao pretendido. Por exemplo, suponha-se que se pretende fazer uma pesquisa por artigos matemáticos cujo conteúdo se foca somente em teoria de grafos. O que acontece é que grande parte da informação disponibilizada após a pesquisa não é relevante para a pergunta inicial e a pesquisa torna-se pouco fiável e incipiente para o utilizador. Nas próximas secções, será explicada de maneira mais pormenorizada os conceitos base e toda a arquitectura que os catálogos científicos devem possuir.

5.2 Modelo Matemático

5.2.1 Catálogos científicos de referência

Para serem eficientes e úteis, os catálogos científicos têm de, não só maximizar a precisão, como também a forma de ordenar os dados após a pesquisa. Maximizar a precisão da pesquisa para garantir resultados mais relevantes para o utilizador, na medida em que os catálogos científicos matemáticos consistem, maioritariamente, em equações, gráficos, tabelas, números dentro de expressões matemáticas e textos, dificultando a procura. Claramente, os utilizadores necessitam de sistemas de pesquisa personalizados, que procurem e recuperem rapidamente a informação matemática que é mais relevante para as suas necessidades.

Um número elevado de sistemas de pesquisa têm sido construidos, mas sem significativos avanços em termos de crescimento e melhoramento, são exemplos disso o sistema de pesquisa NIST DLMF [46, 67, 66] e o sistema Design Science's MathIndex [60].

Actualmente, existem três tipos de sistemas de pesquisa de matemática que foram construidos e/ou receberam atenção e estudos por parte da comunidade científica. O primeiro tipo de sistemas, baseado em campos de pesquisa (metadados), é desenvolvido em base de dados pelos

maiores sistemas de armazenamento de dados matemáticos do mundo, tal como Zentralblatt's ZMath e MathDi [68, 41], a base de dados Jahrbuch [19], AMS MathSciNet [42] e por várias sociedades da matemática existentes no mundo inteiro. Tais sistemas estão direccionados para pesquisa tradicional de documentos.

O segundo tipo de sistemas são semelhantes aos sistemas desenvolvidos por Guidi et al. [27, 28], MOWGLI inserido no projecto Helm [47], e MiZar. Este tipo de sistemas de pesquisa são fortemente especializados, garantem um nível de pesquisa elevado, com garantia de bastante relevância nos documentos devolvidos após a pesquisa, mas não são utilizados pela maioria dos investigadores matemáticos, pois o seu conteúdo é restrito.

O terceiro tipo de sistemas de pesquisa matemático é construído com base em modelos matemáticos de pesquisa, e.g. o DLMF [46, 67, 66] e o sistema Design Sciences MathIndex Web Search. Este tipo de sistemas de pesquisa matemáticos são direccionados para serem maioritariamente utilizados por estudantes, educadores, investigadores científicos e profissionais em Matemática, Fisica e Engenharias. Este tipo de sistemas de pesquisa para serem construidos requerem que se defina uma ordenação à priori da relevância dos dados.

A medição do grau de relevância dos documentos é o principal factor a ter a atenção dos investigadores na construção destes sistemas de pesquisa [7]. Embora muitas métricas de relevância terem sido desenvolvidas e estudadas, muitas são variações de uma distância métrica central, a distância de *if-idf* ("term frequency inverse document frequency"). Essencialmente, esta métrica assume que quanto maior for a frequência relativa de um termo do documento na consulta mais relevante é esse documento. Uma consequência directa desta análise da métrica é que se dois documentos têm o mesmo número de ocorrências de palavras chaves na pesquisa, mas um dos documentos é mais pequeno que o outro, então o documento mais pequeno vai ter um índice de ordenação maior devido ao facto do índice de ordenação ser calculado através do número de ocorrências a dividir pelo tamanho do documento.

Como este método tradicional de cálculo de relevância de um documento não é apropriado para sistemas de pesquisa de documentos matemáticos, nomeadamente, do primeiro tipo anteriormente descrito, apresenta-se nas próximas secções uma proposta de um novo modelo matemático para pesquisa de artigos matemáticos em catálogos científicos.

Por estas e outras razões, torna-se relevante o estudo e desenvolvimento de mecanismos que dada uma necessidade conduzam à obtenção da informação desejada no mais curto espaço

de tempo e que vá ao encontro do que é relevante para o utilizador. Um utilizador que pretenda fazer uma pesquisa avançada sobre documentos relacionados com um determinado tópico da matemática vê-se confrontado com um aglomerado de informação disponível na rede telemática que nem de longe será a mais relevante para si, então conseguir criar e unir num só sistema funcionalidades de pesquisa que permitam satisfazer melhor e mais rapidamente o utilizador é o grande objectivo de todo o trabalho desenvolvido, ultrapassando os resultados dos sistemas comuns com melhor desempenho usando uma função de combinação de pesos do que é mais relevante para o utilizador.

Comece-se por definir abstractamente o conceito de catálogo científico que incorpora os documentos já existentes na MathSciNet e na Zentralblat. Seguidamente apresenta-se uma breve descrição dos códigos `MSC` que indicam os tópicos da matemática englobados maioritariamente no conteúdo dos documentos.

Suponha-se que o sistema poderá ser acedido por um utilizador que pode ser anónimo ou registado. Sendo um utilizador registado terá um "profile"' e mediante esse "profile" a pesquisa é efectuada.

Em concreto, matematicamente, modele-se um catálogo científico como um tuplo $\mathtt{R} = (\mathcal{A}, \mathcal{P}, \mathcal{M}, \mathcal{U}, \mathcal{Q}, \Upsilon_{\mathcal{P}}, \Upsilon_{\mathcal{M}}, \Upsilon_{\mathcal{A}}, \Upsilon_{\mathcal{Q}})$ onde $\mathcal{A}$ é o conjunto de autores, $\mathcal{P}$ conjunto de artigos, $\mathcal{M}$ conjunto de utilizadores registados, $\mathcal{U}$ conjunto de utilizadores comuns (não engloba os registados), $\Upsilon_{\mathcal{P}} : \mathcal{A} \to 2^{\mathcal{P}}$ é o mapa que associa a cada autor o conjunto dos seus artigos, $\Upsilon_{\mathcal{M}} : \mathcal{M} \to \mathcal{U}$ o mapa que identifica cada membro registado como um utilizador, $\Upsilon_{\mathcal{A}} : \mathcal{A} \to \mathcal{U}$ o mapa que identifica cada autor como um utilizador do sistema, $\Upsilon_{\mathcal{Q}} : \mathcal{Q} \to 2^{\mathcal{P}}$ o resultado da consulta (isto é, o processo de consulta ao sistema R com um conjunto de termos $t \in \mathcal{Q}$ e recebe como resultado o conjunto de artigos $\Upsilon_{\mathcal{Q}}(t) \subseteq \mathcal{P}$) onde $\Upsilon_{\mathcal{Q}}$ verifica as seguintes condições:

1. $\mathcal{Q} = \Upsilon_{\mathcal{Q}}^{-1}(2^{\mathcal{P}} \backslash \emptyset)$;
2. $\Upsilon_{\mathcal{Q}}$ é injectiva.

O conjunto $\mathcal{Q}$ poderá conter palavras reservados, como por exemplo operadores de lógica, ou filtros que enriquecem o control da saida de $\Upsilon_{\mathcal{Q}}$.

De forma resumida, o catálogo científico é visto com um tuplo, $\mathtt{R}$, em que são satisfeitas as seguintes condições:

1. Os conjuntos $\mathcal{A}, \mathcal{P}, \mathcal{M}, \mathcal{U}$ e $\mathcal{Q}$ são não vazios;
2. O mapa $\Upsilon_{\mathcal{M}}$ é injectivo, logo pode-se considerar $\mathcal{M} \subseteq \mathcal{U}$;
3. O mapa $\Upsilon_{\mathcal{P}}$ é sobrejectivo e $\mathcal{A} \backslash \Upsilon_{\mathcal{P}}^{-1}(\mathcal{P}) = \emptyset$;
4. Todos os elementos de $\mathcal{P}$ são classificados com códigos `MSC` (pelo menos com o código primário e de um até cinco códigos secundarios);
5. O mapa $\Upsilon_{\mathcal{A}}$ é injectivo, logo pode-se considerar $\mathcal{A} \subseteq \mathcal{U}$.

Pode-se ainda afirmar que `R` é *bem definido* se verifica as condições (1)-(3); *MSC classificado* se verifica a condição (4); *coerente* se verifica a condição (5); e *completo* se verifica todas as condições (1)-(5).
Os conjuntos $\mathcal{A}, \mathcal{P}$ e $\mathcal{M}$ são conjuntos permanentes do catálogo, e $\mathcal{Q}$ é o resultado da pesquisa ao sistema, o conjunto $\mathcal{U}$ é de certa forma um conjunto externo ao sistema e ambíguo. A cada utilizador $u \in \mathcal{U}$ terá de corresponder (temporária ou permanente) uma personagem real ou um conjunto de personagens, através da implementação de alguns mecanismos de autenticação na rede telemática. Esta correspondência e associação poderá ser resolvida de várias maneiras, tem-se a seguir duas das muitas maneiras complexas de implementar essa associação:

- O sistema restringe as suas grandes funcionalidade apenas a membros registados no sistema, logo nesse caso ter-se-á $\mathcal{M} \equiv \mathcal{U}$, onde cada utilizador terá de ser autenticado pelo sistema através de um web login;
- Existe no sistema um tipo de utilizador especial, denominado utilizador anónimo, e para todos a login para autenticação dos utilizadores (incluindo membros registados) é obrigatória. Quando a autenticação dos membros falha ou as tentativas de autenticação forem esgotadas estes são dados como membros não registados no sistema e estes são automaticamente associados a utilizadores anónimos.

Pode-se afirmar ainda que sistemas de pesquisa como o Google e o Scopus não são completos, pois não satisfazem a condição (4), isto é, nem todos os artigos armazenados nestes dois sistemas têm uma classificação MSC com pelo menos um código primário. Por diversas razões a partir de agora consideram-se apenas catálogos científicos completos.

5.2.2 Distância entre códigos MSC

A percepção humana da informação é um processo não linear o que torna bastante difícil propor e descrever numa métrica que defina na perfeição a correlação entre a apreensão e o julgamento que o ser humano faz da informação. O Mathematics Subject Classification (MSC) é um sistema de categorização do conteúdo de informação; ou seja, artigos de matemática, que existe nos sistemas de pesquisa MathSciNet e Zentralblatt. Esta categorização dos documentos tenta fazer a diferença quando um utilizador pretende aceder a um determinado tópico de matemática que se encontra em artigos armazenados em catálogos científicos.
Todos os artigos destes dois catálogos científicos (MathSciNet e Zentralblatt) são categorizados com uma classificação composta por quatro dígitos e uma letra, onde cada um dos dígitos e letra representam um tópico matemático (por exemplo: 11 = "Teoria de números"; $11B$ = "Sequências e conjuntos"; $11B05$ = "Densidade, topologia") e quando um determinado artigo abrange diferentes tópicos da matemática, a classificação é feita através do mais importante de todos os tópicos existentes no mesmo.
Para cada artigo inserido no sistema é-lhe atribuído um código MSC primário para cobrir o tema principal do artigo, e um ou mais códigos MSC secundários para cobrir resultados auxiliares ao tema principal, motivação ou origem e aplicação do problema discutido no artigo, num máximo de 8 códigos.
No entanto, observou-se que esta classificação de artigos pode ser organizada numa árvore de pesquisa, onde todas as categorias são representadas pelos nodos internos da árvore. Todos os artigos são categorizados a partir da árvore de códigos, passando estes documentos a formarem as folhas da árvore. Veja-se por exemplo um artigo que foi catalogado com o seguinte código MSC: 03B40. Na árvore de pesquisa o percurso que leva ao código é o seguinte:

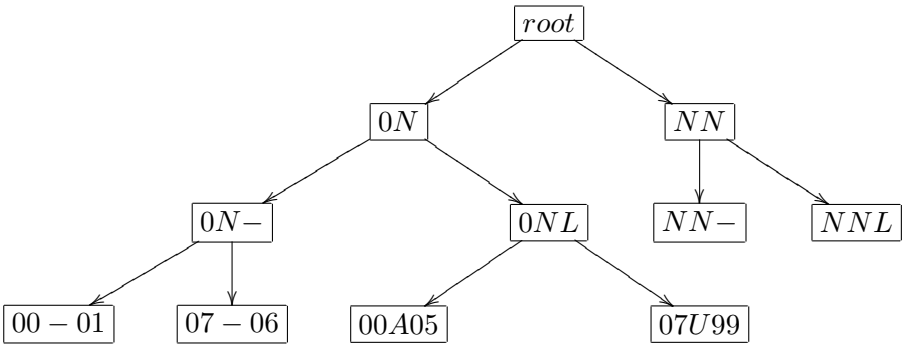
- 03-XX "Mathematical logic And Foundations"
  - 03BXX "General Logic"
    - 03BXX "General Logic"
      - 03B40 "Combinatory logic and lambda-calculus"

Este tipo de estrutura poderá ser útil para a personalização da pesquisa dos artigos através dos seus códigos resultantes da categorização automaticamente feita pela MathSciNet e Zentralblatt. Deste modo, a pesquisa poderá tornar-se mais rápida, eficiente e de maior relevância para um utilizador.

Numa primeira etapa algumas questões se colocam acerca da maneira de integrar este tipo de classificação em árvore num catálogo científico, podendo-se partir mesmo das seguintes observações:

- Se a classificação `MSC` pode ser organizada e representada por uma árvore de pesquisa, então também se poderão calcular todos os caminhos mais curtos entre todas as classificações;
- Sendo possível o cálculo de todos os caminhos mais curtos então pode-se resolver o problema do cálculo do nodo mais profundo comum entre dois códigos da árvore.

Assim sendo, organizando as classificações numa árvore de pesquisa obtem-se a seguinte árvore:



Seja $C = c_1...c_n$ um código `MSC` qualquer. Defina-se a cabeça do código C de tamanho k por

$$H_k(C) = c_1...c_k, 1 \leq k \leq n \tag{5.1}$$

e a cauda do código C de tamanho k por

$$T_k(C) = c_{k+1}...c_n, 1 \leq k \leq n \tag{5.2}$$

Sabendo que o código C tem no máximo comprimento cinco e considerando a estrutura da árvore definida anteriormente, pode-se definir a função h, que representa o nível do nodo mais

profundo comum entre dois códigos por

$$h(C_1, C_2) = \begin{cases} 3 - \Delta & \text{,outros casos} \\ 3 & \text{,} H_2(C_1) \neq H_2(C_2) \end{cases} \tag{5.3}$$

onde $\Delta = \lceil \frac{1}{2} argmax_{1 \leq k \leq n} \{T_k(C_1) \neq T_k(C_2) \wedge H_k(C_1) = H_k(C_2)\} \rceil$ e $\lceil x \rceil$ é o menor inteiro maior que x.

A partir da árvore proposta podem ser calculados os níveis de profundidade entre códigos. Tem-se por objectivo definir uma distância entre dois conjuntos de códigos: um formado pelos códigos que representam os interesses do utilizador e o outro os códigos do catálogo. No entanto, a principal restrição para tal distância é o facto de ambos os conjuntos ignorarem a estrutura específica da árvore hierárquica de códigos `MSC` definida anteriormente. Para contornar este problema ir-se-á introduzir uma métrica na estrutura de códigos `MSC`, baseada em trabalhos anteriores sobre métricas em palavras. Por exemplo, para calcular a similaridade entre palavras, usando o conhecimento da hierarquia semântica definido por J. Miller [4] a fórmula desenvolvida para o cálculo da distância entre dois nodos da árvore pode ser dada pela expressão

$$\mu(h, L) = e^{-\alpha L} \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} = e^{-\alpha L} tanh(\beta h), \tag{5.4}$$

onde os parâmetros α e β são definidos como 0.2 e 0.6 respectivamente, h é o nodo mais profundo comum a dois nodos, e L é o comprimento do caminho semântico entre duas palavras. Para a estrutura `MSC` estudada neste trabalho e tendo em conta a função descrita em (5.3) e definindo $L(c_1, c_2) = 6 - 2h(c_1, c_2)$ como o caminho mais curto entre dois códigos quaisquer, tem-se que a função

$$P(c_1, c_2) \quad = \quad \frac{\mu\left(h(c_1, c_2), L(c_1, c_2)\right)}{\mu(3, 0)}.$$

define uma função de proximidade normalizada entre dois códigos, $c_1, c_2 \in$ `MSC` (tomando valores entre 0 e 1).

Definição 26 *Defina-se por função distância D em relação ao espaço métrico (X, D) a função que satisfaz as seguintes condições, para todo o $x, y, z \in X$:*

(1) $D(x, y) \geq 0$;

(2) $D(x, x) = 0$;

(3) $x \neq y$ *implica que* $D(x,y) > 0$;

(4) $D(x,y) = D(y,x)$;

(5) $D(x,z) \leq D(x,y) + D(y,z)$.

Uma função distância no conjunto de códigos MSC é descrita pelo seguinte lema

Lema 27 *A função*

$$D_{MSC}(c_1,c_2) = 1 - \frac{sinh(0.6h(c_1,c_2)+1.8) + sinh(0.6h(c_1,c_2)-1.8)}{cosh(0.6h(c_1,c_2)+1.8) - cosh(0.6h(c_1,c_2)-1.8)} e^{(0.4h(c_1,c_2)-1.2)}$$

é uma função distância (normalizada) no conjunto de códigos MSC.

Prova Para provar que a distância $D_{\texttt{MSC}}$ define uma distância em MSC ter-se-á de provar as condições descritas na definição 26.

1. $D_{\texttt{MSC}}(c_1,c_2) = 0 \iff c_1 = c_2$.
   $D_{\texttt{MSC}}(c_1,c_2) = 1 - P(c_1,c_2) = 1 - \frac{\mu(h(c_1,c_2),L(c_1,c_2))}{\mu(3,0)} = 0 \iff 1 = \frac{\mu(h(c_1,c_2),L(c_1,c_2))}{\mu(3,0)} \iff \mu(3,0) = \mu(h(c_1,c_2),L(c_1,c_2)) \iff h(c_1,c_2) = 3 \wedge L(c_1,c_2) = 0 \iff h(c_1,c_2) = 3 \wedge 6 - 2h(c_1,c_2) = 0 \iff c_1 = c_2$

2. $D_{\texttt{MSC}}(c_1,c_2) \geq 0, \forall c_1, c_2$
   $D_{\texttt{MSC}}(c_1,c_2) = 1 - \frac{\mu(h(c_1,c_2),L(c_1,c_2))}{\mu(3,0)} \geq 0 \iff 1 \geq \frac{\mu(h(c_1,c_2),L(c_1,c_2))}{\mu(3,0)} \iff \mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2)) \iff 1 \geq \mu(h(c_1,c_2),L(c_1,c_2)) \iff 1 \geq P(c_1,c_2) \iff 1 - P(c_1,c_2) \geq 0$

3. $D_{\texttt{MSC}}(c_1,c_2) + D_{\texttt{MSC}}(c_2,c_3) \geq D_{\texttt{MSC}}(c_1,c_3) \iff \mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2)) + \mu(h(c_2,c_3),L(c_2,c_3)) - \mu(h(c_1,c_3),L(c_1,c_3))$
   Suponham-se os dois casos seguinte :

   (a) $c_1 = c_2$:
   $\mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2)) + \mu(h(c_2,c_3),L(c_2,c_3)) - \mu(h(c_2,c_3),L(c_2,c_3)) \iff \mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2))$

   (b) $c_1 \neq c_2$:

i. $c_1 = c_3$:
$\mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2))+\mu(h(c_2,c_1),L(c_2,c_1))-\mu(h(c_1,c_1),L(c_1,c_1)) \Longleftrightarrow$
$\mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2))$

ii. $c_1 \neq c_3$:
$\mu(3,0) \geq \mu(h(c_1,c_2),L(c_1,c_2))+\mu(h(c_2,c_3),L(c_2,c_3))-\mu(h(c_1,c_3),L(c_1,c_3))$,
$h(c_1,c_2) \neq 3$,$h(c_2,c_3) \neq 3$ and $h(c_1,c_3) \neq 3$ então tem-se que
$\mu(h(c_1,c_3),L(c_1,c_3)) < \mu(3,0) \Longleftrightarrow -\mu(h(c_1,c_3),L(c_1,c_3)) > -\mu(3,0)$então
$\mu(3,0) > \mu(h(c_1,c_2),L(c_1,c_2)) + \mu(h(c_2,c_3),L(c_2,c_3)) - \mu(3,0)$.

□

5.2.3 Espaço vectorial livre gerado pelos códigos `MSC`

As razões de definir um espaço vectorial livre gerado por códigos `MSC` são diversas. Uma dessas razões é o facto de ser fundamental na obtenção de uma representação útil dos documentos existentes no catálogo em termos de códigos `MSC`, a possibilidade de realizar operações com esses documentos (soma, subtracção e multiplicação de um valor constante por cada documento).

Um espaço vectorial livre relativamente a determinado campo é um conjunto finito de vectores fechado em relação à adição e multiplicação por um escalar, onde os escalares são elementos do campo escolhido. Também é possível definir um espaço vectorial livre como um conjunto de vectores capaz de ser substituído por um outro conjunto qualquer.

Um espaço vectorial livre real, V_x gerado por um conjunto X é caracterizado como sendo um espaço vectorial em $\mathbb{R}$ cujo conjunto de vectores é definido por

$$\{\phi : X \to \mathbb{R} : \phi^{-1}(\mathbb{R}\backslash\{0\}) \text{ é finito}\}. \tag{5.5}$$

A estrutura do espaço vectorial resulta directamente das propriedades gerais das funções, [24]. Para cada $W \in X$ define-se a função $\aleph_w \in V_X$ por

$$\aleph_w(x) = \begin{cases} 1 & x = w \\ 0 & \text{caso contrário.} \end{cases} \tag{5.6}$$

[1] MR2189889 Higham, Desmond J. Google PageRank as mean playing time for pinball on the reverse web. Appl. Math. Lett. 18 (2005), no. 12, 1359–1362. 68U35 (60J10 60J20)

(**Código MSC primário**: 68U35; **Lista de códigos MSC secundários**: 60J10 60J20.)

Figura 5.1: Exemplo de um artigo referenciado na MathSciNet

O conjunto de funções $\{\aleph_w\}_{w\in X}$ forma uma base linearmente independente para V_X, isto é, qualquer conjunto ϕ de V é uma combinação linear finita, $\phi(\cdot) = \sum_{i=1}^{n} \phi(x_i)\aleph_{x_i}(\cdot)$, onde $S_\phi = \{x_1, \ldots, x_n\}$ é o conjunto de pontos onde ϕ é não nulo. O mapa $h : X \to V_X$, $w \mapsto \aleph_w$ resulta numa bijeção entre X e a base de V_X.

O mapa h é universal no sentido que para qualquer $\psi : X \to W$ arbitrário, com W vector espaço livre, existe um único mapa $\bar{\psi} : V_X \to W$ tal que $\psi = \bar{\psi} \circ h$.

Assim sendo, denotando-se por `MSC` o conjunto finito de códigos da classificação matemática descrita na secção anterior, por N a sua cardinalidade e por $\mathbb{R}_{\texttt{MSC}}$ o espaço vectorial livre gerado a partir de `MSC`.

5.2.4 Documentos como elementos de $\mathbb{R}_{\texttt{MSC}}$

Um item importante que os autores de um artigo matemático devem cuidar na sua submissão de artigos para posterior publicação é a atribuição dos códigos `MSC`. Tais códigos são importantes para a indexação dos documentos no catálogo cientifico. Estes códigos podem ser organizados em estruturas de forma hierárquica, de modo a facilitar a pesquisa e recuperação da informação. Na especificação `MSC`, descrita por M. Bouklit [39], um documento é classificado por um código primário e por um conjunto de códigos secundários, como se observa na figura 5.1. Defina-se deste modo, por $prm : \mathcal{A} \to \texttt{MSC}$ e $sec : \mathcal{A} \to 2^{\texttt{MSC}}$ o código primário e o conjunto de códigos secundários, respectivamente. Fixe-se um parâmetro $\epsilon_p \in [0, 1]$ que representa o peso dos códigos primários num documento. Qualquer documento do sistema pode, então, ser representado como um elemento de $\mathbb{R}_{\texttt{MSC}}$ através da construção do mapa

$\Xi : \mathcal{P} \to \mathbb{R}_{\texttt{MSC}}$ definido por

$$P \mapsto \epsilon_p \aleph_{prm(P)} + \frac{1}{|sec(P)|} \sum_{c \in sec(P)} (1 - \epsilon_p)\aleph_c. \tag{5.7}$$

Considere-se o seguinte exemplo que demonstra a aplicação da construção do mapa Ξ para (5.1), onde $\epsilon_p = 1/2$ e P_1, P_2 são os documentos.

$$\Xi(P_1) = \frac{1}{2}\aleph_{68U35} + \frac{1}{4}\aleph_{60J10} + \frac{1}{4}\aleph_{60J20}, \tag{5.8}$$

$$\Xi(P_2) = \frac{1}{2}\aleph_{68U35} + \frac{1}{6}\aleph_{60J10} + \frac{1}{6}\aleph_{60J20} + \frac{1}{6}\aleph_{60J23}, \tag{5.9}$$

$$3\Xi(P_1) - \Xi(P_2) = \aleph_{68U35} + \frac{7}{12}\aleph_{60J10} + \frac{7}{12}\aleph_{60J20} - \frac{1}{6}\aleph_{60J23}. \tag{5.10}$$

5.2.5 Autores como elementos de $\mathbb{R}_{\texttt{MSC}}$

Qualquer autor pode ter publicado um ou mais artigos no sistema `R`. Defina-se desta forma, por $P_A \equiv \Upsilon_{\mathcal{P}}(A)$ o conjunto de todos os artigos publicados pelo autor $A \in \mathcal{A}$.
Sabe-se da secção anterior que cada artigo $P \in \mathcal{P}$ do autor A é um elemento de $\mathbb{R}_{\texttt{MSC}}$, então um autor A pode também ser visto como um elemento de $\mathbb{R}_{\texttt{MSC}}$ através da construção do mapa $\nu : \mathcal{A} \to \mathbb{R}_{\texttt{MSC}}$ dado por

$$\nu(A) = \sum_{P \in P_A} \Xi(P) = \sum_{P \in P_A} [\epsilon_p \aleph_{prm(P)} + \frac{1}{|sec(P)|} \sum_{c \in sec(P)} (1 - \epsilon_p)\aleph_c]. \tag{5.11}$$

Normalizando a expressão dada tem-se que

$$\nu_N(A) = \frac{1}{|P_A|}\nu(A) \sim P_{virtual}, \tag{5.12}$$

ou seja, para cada autor $A \in \mathcal{A}$, $\nu_N(A)$ corresponderá a um vector descritor de um artigo virtual com todas as entradas definindo os pesos de cada código `MSC` que todos os seus artigos têm a classificá-los. Este resultado será útil para a classificação automática de autores que será discutido no próximo capítulo.

5.3 Uma distância em $\mathbb{R}_{\texttt{MSC}}$

Considere-se $\prec$ uma ordem total em `MSC` que define uma sequência crescente de elementos

$\texttt{MSC}^{\prec} = \{w_1, \ldots, w_N\}$, onde $w_1 \prec \cdots \prec w_N$. A mesma permite a construção do vector representativo de $\mathbb{R}_{\texttt{MSC}}$ e do mapa $\varphi : \mathbb{R}_{\texttt{MSC}} \to \mathbb{R}^N$ definido por

$$v = \sum_{w \in \texttt{MSC}} v_w \aleph_w \mapsto (v_{w_1}, v_{w_2}, \ldots, v_{w_N}) \tag{5.13}$$

onde o mapa φ é uma bijecção. Defina-se por descritor MSC qualquer que seja o vector pertencente à imagem de φ, e MSC descritor de um artigo $A \in \mathcal{P}$ a qualquer que seja o vector $\varphi(\Xi(A)) \in \mathbb{R}^n$. É fácil de reconhecer de imediato que todos os descritores MSC pertencem ao cone $\mathbb{R}^N_+$. No entanto, um problema se pode colocar, é como se poderá definir uma distância em $\mathbb{R}_{\texttt{MSC}}$?

Têm de ser avaliados dois tipos de dificuldades acrescidas, o facto da existência dos coeficientes dos descritores MSC e as distâncias entre os respectivos códigos MSC referida na secção anterior. Em particular, como medir a distância entre $-2\aleph_w(68U35)+1/6\ \aleph_w(60J10)$ e $1/3\aleph_w(60J20)-\sqrt{2}\ \aleph_w(60J23)$? Aplicar a definição de distâncias métricas em $\mathbb{R}^n$ não será a abordagem mais adequada, nem mesmo aplicar produto de métricas.

Para introduzir uma função distância em $\mathbb{R}_{\texttt{MSC}}$, que considere os descritores MSC e, também, a própria distância entre códigos, iremo-nos basear na conhecida distância de Mahalanobis.

A distância de Mahalanobis, introduzida por P.C. Mahalanobis em 1936, é uma distância da estatística baseada nas correlações entre as variáveis em estudo. É útil para determinar a similaridade entre um conjunto modelo desconhecido e um previamente conhecido.

Este tipo de distância difere da distância euclidiana na medida em que toma como relevante as correlações entre o conjunto de dados, não dependendo da escala de medição. Assim sendo, formalmente a distância de Mahalanobis entre dois conjuntos de pontos $Y = (y_1, y_2, ..., y_n)^T$ e $X = (\ x_1, x_2, ..., x_n)^T$ é definida por

$$D_M(X, Y) = \sqrt{(X - Y)^T S^{-1} (X - Y)}. \tag{5.14}$$

A distância em $\mathbb{R}_{\texttt{MSC}}$ será do tipo da distância de Mahalanobis.

Teorema 5.3.1 *Sejam $X, Y \in \mathbb{R}_{\texttt{MSC}}$. A função*

$$D_M(X, Y) = \sqrt{(X - Y)^T M^{-1} (X - Y)} \tag{5.15}$$

é uma função distância, onde a matriz M é definida por

$$M = \{a_{ij}\}_{i,j=1}^N \quad com\ a_{ij} = 1 - D_{\texttt{MSC}}(c_i, c_j) \tag{5.16}$$

e $c_i \in MSC^{\prec}$.

Prova Seja M uma matriz real, simétrica e definida positiva cuja diagonal principal tem entradas todas iguais a um. M admite inversa e é diagonalizável, i. e., semelhante a uma matriz diagonal. Pode-se então afirmar que existe uma matriz invertível P tal que $P^{-1}MP = D$, onde P é a matriz diagonalizante de M. Como D é invertível pode-se afirmar que

$$D^{-1} = (P^{-1}MP)^{-1} = P^{-1}M^{-1}P \tag{5.17}$$

onde P é a matriz diagonalizante de M^{-1} e $M^{-1} = PD^{-1}P^{-1}$. Como $P^T = P^{-1}$, i. é, P é ortogonal tem-se então que $M^{-1} = PD^{-1}P^T$.
Por substituição em (5.15) obtém-se

$$\begin{aligned} D_M(X,Y) &= \sqrt{(X-Y)^T PD^{-1}P^T(X-Y)} \\ &= \sqrt{(P^T(X-Y))^T D^{-1}P^T(X-Y)} \\ &= \sqrt{(P^TX - P^TY)^T D^{-1}(P^TX - P^TY)} \end{aligned} \tag{5.18}$$

Denotando $\overline{X} = P^TX$ e $\overline{Y} = P^TY$ resulta

$$D_M(X,Y) = \sqrt{(\overline{X}-\overline{Y})^T D^{-1}(\overline{X}-\overline{Y})} \tag{5.19}$$

Fazendo $X^* = \frac{\overline{X}}{\sqrt{d_i}}$ e $Y^* = \frac{\overline{Y}}{\sqrt{d_i}}$ a distância definida em (5.15) passa a ser da forma

$$D_M(X,Y) = d_E(X^*,Y^*) \tag{5.20}$$

onde d_E representa a distância euclidiana. Fica assim provado que a função D_M define uma distância. Portanto, a distância proposta é a distância Euclidiana de pontos obtidos por transformação dos pontos iníciais. □

Note-se que uma matriz quadrada A diz-se diagonalizável se, em relação a certa base de um espaço vectorial, é uma matriz de um endomorfismo diagonalizável, i. é., se for semelhante a uma matriz diagonal. A uma matriz invertível P tal que $P^{-1}AP$ é uma matriz diagonal chama-se matriz diagonalizante de A. Para além disso, se A é uma matriz simétrica, i.e., $A = A^T$ então A é diagonalizável.
No capítulo seguinte ir-se-á introduzir diversas funcionalidades para um catálogo científico baseadas na função de distância proposta.

Capítulo 6

Novas Funcionalidas de um Catálogo Científico

Pretende-se que cada utilizador do catálogo científico indique e seleccione os seus interesses ou tópicos matemáticos mais relevantes traduzidos posteriormente por códigos `MSC`, seguidamente denominado por o "profile" do utilizador.
Interpretando os tópicos de interesse do utilizador através dos códigos `MSC`, poderão ser calculadas todas as distâncias entre os códigos do "profile" do utilizador e os códigos existentes no sistema, e posteriormente ser realizada uma ordenação de artigos de relevância para uma consulta efectuada. Seja U um utilizador qualquer do sistema, define-se por profile de U

$$Profile(U) = \{v_N(A) + (1-\xi)v_p : \xi \in [0,1]\} \tag{6.1}$$

onde ξ é um coeficiente de escala, v_p é um vector descritor `MSC` que codifica as preferências do utilizador e onde

$$A = \begin{cases} 0, & \text{U não é autor;} \\ \Upsilon_A^{-1}(U), & \text{caso contrário.} \end{cases} \tag{6.2}$$

6.1 Ordenação

Uma vez definida D_M (teorema 5.14) a execução do processo de ordenação dos dados por

relevância pode ser descrito e calculado computacionalmente através do algoritmo seguidamente apresentado.

Algoritmo : Pesquisa personalizada

Entrada: U : Utilizador; Q : consulta.

1: Obter o vector descritor dos interesses do utilizador e formar o profile de U, P_U.
2: Enviar Q para pesquisar no catálogo;
3: Res = Vector de URL's retornado pelo catálogo;
4: V=[ ];
5: **for** $j = 1$ to size(Res) **do**
6: $V[j] = D_M(Res[j], P_U[j])$
7: **end for**
8: Ordenar o vector Res em função dos valores no vector V.

Saida: Res :Vector de URL's, ordenada segundo a preferência do utilizador U.

6.2 Agrupamento de Metadados

A grande parte dos sistemas de informação das mais diversas áreas do conhecimento têm armazenado e produzido uma grande quantidade de dados, gerados pela facilidade de troca, armazenamento e disponibilidade de informação existente hoje em dia na Internet.
No entanto, existe a necessidade de agrupar e classificar esses dados em conjuntos de dados de menor dimensão e com características semelhantes, implicando a necessidade da utilização de métodos de classificação e análise de dados, designados por métodos de agrupamento. Agrupamento é o processo de divisão de um conjunto de dados em subgrupos de objectos similares, onde cada grupo consiste num conjunto de objectos que são similares entre eles mas diferentes dos restantes grupos [8].
Os métodos de agrupamento são usualmente divididos em dois grupos: métodos hierárquicos

ou particionados. Por sua vez, os métodos hierárquicos de agrupamento são subdivididos em aglomerativos ou divisivos [8]. A base dos métodos de agrupamento hierárquicos inclui a fórmula de Lance-Williamns [16], ideia originária do método de agrupamento. Enquanto que os algoritmos hierárquicos constroem os grupos, "clusters", de forma gradual, os algoritmos particionais constroem os grupos de forma directa. Durante a execução do algoritmo particional, tanto se tenta descobrir os grupos iteractivamente através da re-colocação de pontos entre os conjuntos ou tentanto identificar grupos em áreas com grande densidade de dados.
Nos métodos hierárquicos aglomerativos os dados são inicialmente distribuidos de modo que cada exemplar represente um grupo e, então, esses grupos são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplares pertençam a apenas um grupo. Na abordagem dos métodos hierarquicos divisivos, o processo inicia-se com apenas um clusters, contendo todos os dados, e segue dividindo-se recursivamente, segundo alguma métrica, até que alcance algum critério de paragem.
Uma das principais caracteísticas dos métodos de agrupamento hierárquicos é a flexibilidade na análise dos grupos, nos diferentes níveis de agrupamento, o que naturalmente sugere um refinamento na análise dos padrões neles descritos [8].
Considerando o catálogo cientifico $\mathtt{R} = (\mathcal{A}, \mathcal{P}, \mathcal{M}, \mathcal{U}, \mathcal{Q}, \Upsilon_{\mathcal{P}}, \Upsilon_{\mathcal{M}}, \Upsilon_{\mathcal{A}}, \Upsilon_{\mathcal{Q}})$, uma funcionalidade que este sistema poderá apresentar ao utilizador é o agrupamento de artigos pela existência de similaridade ou não entre eles. Desta forma, existe a necessidade da utilização de métodos e análise de aglomeração no sistema para fazer o agrupamento de dados.
No capítulo anterior foi definida uma nova distância, a distância D_M em $\mathbb{R}_{\mathtt{MSC}}$, definida em 5.14. Com essa distância poder-se-á construir uma matriz quadrada, A, de dimensão igual ao número de artigos existentes em $\mathtt{R}$, onde A_{ij} representa o grau de similaridade entre os vectores v_i e v_j, com $v_i, v_j \in \mathbb{R}_{\mathtt{MSC}}$, i. e., $A_{ij} = 1 - D_M(v_i, v_j)$. A partir do grau de similaridade entre os vectores descritores facilmente se poderá concluir acerca do grau de semelhança entre os artigos do sistema.
O agrupamento de artigos em conjuntos de menor dimensão, descritos como clusters, cujo grau de similariedade entre os elementos de cada conjunto é grande e entre elementos de conjuntos distintos ser pequena ou mesmo inexistente pode ser o ponto de partida deste processo. Assim sendo, e depois de calcular a matriz A, o próximo passo é determinar um coeficiente de associação entre os elementos de A. A fórmula para o cálculo do coeficiente de associação

entre os artigos é definida do seguinte modo

$$B_{ij} = \frac{A_{ij}^2}{A_{ii}A_{jj}} \tag{6.3}$$

obtendo-se de imediato uma nova matriz, B, com dimensão igual ao número total de artigos do sistema R. A matriz B é definida a partir dos valores de coeficiente de associação entre todos os elementos de R. Deste modo, pode-se afirmar que:

$$B_{ij} \quad \forall i \neq j; i, j = 1, ..., N, \tag{6.4}$$

e

$$B_{ii} = 1; \quad \forall i = 1, ..., N. \tag{6.5}$$

O próximo passo será construir "clusters" a partir da matriz de associações B. Estes clusters são grupos de dados de menores dimensões. Para isso pode-se utilizar um dos métodos de "clustering" existentes e referidos anteriormente. Um exemplo do algoritmo de agrupamento apresenta as seguintes etapas:

1. Inicialmente, cada entrada da matriz B é representativa de um cluster:
   B_{ij} representa o cluster formado por P_i e P_j com $P_i, P_j \in \mathcal{P}$ e $i \neq j$.

2. Do conjunto de artigos são escolhidos aleatoriamente, dos que tiverem menor valor de coeficiente de associação, pares de artigos, dois a dois de maneira a serem examinados sequencialmente para posterior construção de clusters.

3. Se o par de artigos já pertence ao mesmo cluster então a associação entre eles é definida como uma associação interna, mantendo-se inalterado;

4. Caso pertençam a clusters diferentes o algoritmo tenta agrupa-los num só cluster.

Este processo terminará quando todos os artigos estiverem agrupados no menor número possivel de clusters. No entanto, o utilizador poderá modificar os parâmetros de construir os grupos ("clusters").

6.3 Vizinhança de um Artigo

Considerem-se dois artigos P_i e P_j com $P_i, P_j \in \mathcal{P}$ e $i \neq j$, e calcule-se os vectores descritores

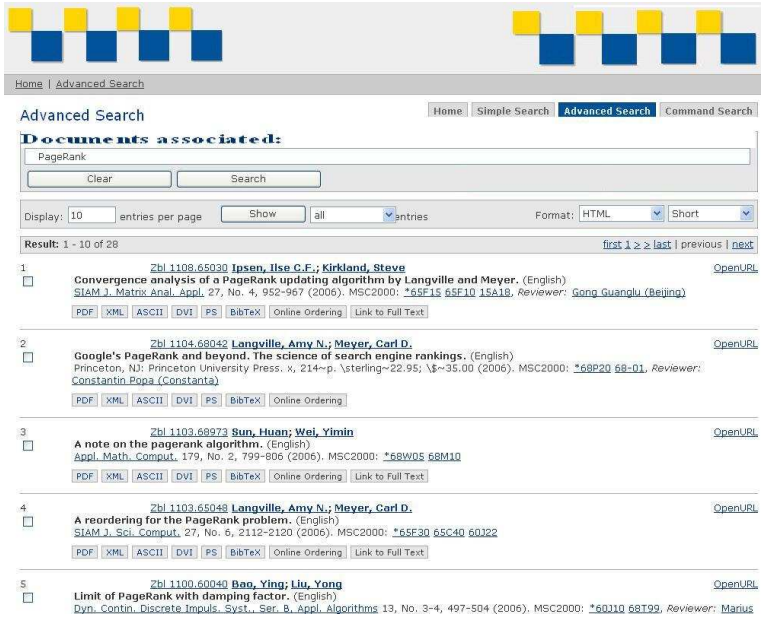


Figura 6.1: Protótipo de interface para pesquisa de artigos na vizinhança de outro artigo.

$v_i \equiv \Xi(P_i)$ e $v_j \equiv \Xi(P_j)$, respectivamente. Defina-se então a distância $D_M(v_i, v_j)$; quanto mais próximo de zero estiver o valor desta distância, mais perto em termos de conteúdo estão os dois artigos. De outra forma, pode-se afirmar que o artigo P_i pertence à vizinhança do artigo P_j e vice-versa. De forma análoga à descrita, pode-se recursivamente, construir o conjunto de vizinhanças de um qualquer artigo do sistema. Esta funcionalidade poderá ser implementada no catálogo científico descrito, permitindo ao utilizador pesquisar sobre outros artigos publicados no sistema com conteúdos semelhantes a um artigo em particular. Observe-se a imagem 6.1 que representa um protótipo de interface possível para pesquisa no catálogo científico de artigos na vizinhança de outro artigo.

Seja $N_\epsilon(P)$ o conjunto de artigos numa vizinhança $\epsilon > 0$ de P, i. é.

$$N_\epsilon(P) = \{\overline{P} \in P : d(\Xi(P), \Xi(\overline{P})) < \epsilon\} \tag{6.6}$$

Seja $\delta_D > 0$ um valor pré definido do número máximo de artigos que se pretende visualizar e $\delta_\epsilon > 0$ a maior distância que se pretende considerar então o algoritmo que determina os artigos a mostrar em "Ver artigos relacionados" da figura 6.1 a P é

Algoritmo: Vizinhança de um Artigo

Entrada: P, δ_D, $\delta_\epsilon > 0$

1: N=n; V=[ ];
2: **for** $i = 1$ to N **do**
3: $R = i\ \frac{\delta_\epsilon}{N}$
4: **if** $|N_R(P) + V| \leq \delta_\epsilon$ **then**
5: $V = V \cup N_R(P)$
6: **else if** $|V| > 0$ **then**
7: return(V);
8: **else**
9: $V = N_R(P)$;
10: return(V);
11: **end if**
12: **end for**

Saida: V : lista ordenada de artigos.

6.4 Vizinhança de um Autor/Utilizador

Considerem-se dois autores A_i e A_j, com $A_i, A_j \in \mathcal{A}$ e $i \neq j$. Calcule-se os vectores descritores $a_i \equiv v_A(A_i)$ e $a_j \equiv v_A(A_j)$ respectivamente. Assim, é possível calcular a distância $D_M(a_i, a_j)$, que indica o nível de proximidade entre os dois autores. De forma generalizada, a partir do algoritmo descrito na secção anterior e com algumas variações, pode-se calcular a vizinhança de um autor/utilizador.

Seja $N_\epsilon(A)$ o conjunto de autores/utilizadores numa vizinhança $\epsilon > 0$ de A, i. é.

$$N_\epsilon(A) = \{\overline{A} \in A : d(\Xi(A), \Xi(\overline{A})) < \epsilon\} \tag{6.7}$$

Seja $\delta_D > 0$ um valor pré definido do número máximo de autores/utilizadores que se pretende visualizar e $\delta_\epsilon > 0$ a maior distância que se pretende considerar. Este algoritmo passa a estar

definido da seguinte forma:

Algoritmo: Vizinhança de um Autor/Utilizador

Entrada: $A,\ \delta_D,\ \delta_\epsilon > 0$

1: N=n; W=[ ];
2: **for** $i = 1$ to N **do**
3: $R = i\ \frac{\delta_\epsilon}{N}$;
4: **if** $|N_R(A) + W| \leq \delta_\epsilon$ **then**
5: $W = W \cup N_R(A)$
6: **else if** $|W| > 0$ **then**
7: return(W);
8: **else**
9: $W = N_R(A)$;
10: return(W);
11: **end if**
12: **end for**

Saida: W: lista ordenada de autores/utilizadores.

6.5 Classificação Automática de Autores

De $v_N(A)$ pretende-se extrair um conjunto bem definido de códigos `MSC`, i. e., um código primário e no máximo $b \leq 8 \in \mathbb{N}$ códigos secundários. O vector descritor do autor definirá de forma percentual em que códigos `MSC` este tem mais artigos escritos, indicando indirectamente os tópicos de especialização do autor.

Com esta ideia subjacente, fazer uma classificação automática de autores poderá ser uma nova funcionalidade a adicionar ao sistema. Com esta classificação pode-se ter maior noção das áreas da matemática de interesse do autor e das suas publicações.

Considere-se então, o vector v, normalizado de $\mathbb{R}_{\texttt{MSC}}$ associado ao autor A, isto é,

$$v = \nu_N(A). \tag{6.8}$$

Considere-se $i^* \in N : v_{i^*} > v_j \ \forall j$ pode-se então garantir que v_{i^*} é a entrada do vector v com valor máximo. Logo, este peso identifica o código que o autor mais utiliza nos seus artigos e consequentemente poderá indicar a sua área principal de trabalho. As próximas b entradas no vector com valores maiores irão representar os códigos secundários. Com esta classificação automática de vectores pode-se ter a percepção das áreas de especialização dos autores.

6.6 Notícias Personalizadas

Notícias personalizadas ou RSS Feeds é um mecanismo/serviço passível de implementar num catálogo cientifico utilizando o modelo matemático proposto. Caracteriza-se por ser um mecanismo que permite distribuir o conteúdo do sistema de forma padronizada e de maneira a que esse conteúdo seja visualizado e divulgado aos utilizadores registados no sistema.
Suponha-se que foi publicado no sistema um novo artigo. Existindo para cada utilizador um profile de interesses e sabendo que esse profile de interesses é constituido por códigos `MSC` o sistema pode de imediato ditar se o novo artigo publicado é do interesse ou não do utilizador, nomeadamente, se $d(\Xi(P), v(Profile(U)) < \epsilon$. Desta forma, é-lhe enviada uma notícia personalizada com acesso imediato ao artigo.
Este método permite que o sistema esteja mais próximo do utilizador e que este tenha acesso mais rápido à informação que é mais relevante. Com este serviço o acesso á informação publicada torna-se mais rápido e mais fácil, isto porque, se por um lado os investigadores publicam artigos em revistas muito distintas, por outro o número de revistas internacionais é elevado, logo é muito díficil a um investigador conseguir acompanhar a publicação de novos artigos de seu interesse.

Capítulo 7

Conclusão e Trabalhos Futuros

7.1 Conclusão

Para dar ínicio a trabalhos científicos é necessário formar e estruturar uma base de conhecimento. Para tal, estudou-se detalhadamente o que existia em termos de conteúdo científico na área em questão. Após a compilação da informação, aprofundaram-se conhecimentos, estudou-se o que existia publicado e criou-se um modelo matemático possível de propor novas soluções para pesquisa em catálogos científicos na área da matemática.

A solução proposta é um modelo matemático a implementar futuramente num sistema de catalogação de artigos de matemática como por exemplo a American Mathematical Society. Este modelo permitirá que o sistema mostre ao utilizador, anónimo ou registrado, de forma organizada numa topologia hierárquica, os documentos mais relevantes de acordo com a sua pesquisa. Para isso, formula-se um mecanismo que calcula a afinidade entre todos os artigos de acordo com uma distância criada a partir dos códigos `MSC` que os mesmos contêm. Para além disso, o sistema poderá englobar um conjunto de novas funcionalidades: permitirá, não obstante a comum pesquisa básica de artigos, que o utilizador registado tenha acesso a notícias personalizadas automaticamente e actualizadas diariamente. Essas notícias chegarão ao

conhecimento do utilizador dando-lhe conta que um novo artigo de relevância para si foi publicado no sistema. Um utilizador/autor poderá ter acesso a uma lista de autores registrados no sistema que têm interesses na mesma área de investigação. Este trabalho é a base de um projecto em desenvolvimento em parceria com a American Mathematical Society.

7.2 Trabalhos Futuros

Para realizar este trabalho de investigação, definiu-se um plano de actividades que não se esgota com o finalizar desta tese. São definidas seguidamente um conjunto de ideias decorrentes da pesquisa de bibliografia e trabalhos relacionados com o tema, que a curto prazo permitirá uma continuação do estudo na área em questão. Ir-se-á salientar brevemente apenas uma dessas ideias.

No universo complexo dos catálogos científicos existem muitos intervenientes. Um desses intervenientes é o avaliador ("reviewer" ou "referee"). Quando um artigo é submetido para publicação tem de satisfazer um conjunto de características para ser aceite. A avaliação das mesmas é feita por avaliadores escolhidos de acordo com certos factores. Os avalidadores têm de ser autores com artigos publicados na área, i. e., têm de ser autores com resultados "relevantes" na área a que o artigo a avaliar se refere, e ter os requisitos intelectuais e críticos para realizar uma boa avaliação. Mediante estes factores, e com a ajuda da percepção humana, é feita a escolha dos avaliadores dos artigos que são diariamente submetidos aos jornais e revistas científicas. Minimizar o impacto destes factores e a aleatoriedade da percepção humana na escolha do avaliador de certo artigo seria fundamental para a evolução dos catálogos científicos. Deste modo, permitir que um sistema faça uma proposta automática do avaliador para determinado artigo é um dos trabalhos que esta dissertação se propõe realizar futuramente. A solução passa por encontrar um modelo de optimização linear que garanta uma escolha óptima do avaliador mediante os factores descritos, baseados nas ferramentas matemáticas introduzidas. Sendo um problema multi-objectivo, a utilização do DEA ("Data Enveloping Analysis") será fundamental para a resolução deste problema em aberto.

Bibliografia

[1] A. Tomking A. Arasu, J. Novak and J.Tomlin, *Pagerank computacion and the struture of the web: Experiments and algorithms*, Behavior Res. Methods Instruments and Computers (1991), 229–236.

[2] J. C. Miller A. Farahat, T. Lofaro and L. A. Ward, *Existence and uniquenness of ranking vectors for linear link analysis*, SIAM (1998), 1–20.

[3] J. C. Miller G. Rae A. Farahat, T. Lofaro and L. A. Ward, *Authority ranking from hits, pagerank and salsa: Existence, uniqueness and effect of initialization*, ACM SIGIR 2001 (2005), 135–161.

[4] A. B. Anjo, *Optimização combinatória*, Dep.Matemática (2005), 13–20.

[5] K. Avrachenkov and N. Litvak, *Decomposition of the google pagerank and optimal linking strategy*, ISSN 0169-2690 (2004), 944–945.

[6] L. G. Robert B. M. Leiner, V. G. Cerf, *A brief history of the internet*, 2003.

[7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Reading, MA (1999), Addison–Wesley.

[8] P. Berkhin, *Survey of clustering data mining techniques.*

[9] K. Bharat and M. R. Henzinger, *Improved algorithms for topic distillation in hyperlinked enviroments*, SIGIR (1998), 104–111.

[10] J. Borbinha, *Digital libraries and the reborn of printed journals*, pp. 97–108.

[11] T. Bouche, *Digitisation and matadata overview*, 2005.

[12] H. Zha C. Ding, X. He and H. Simon, *Pagerank, hits and unified framework for link analysis*, SIGIR (2002), 353–354.

[13] G. H. Golub C. P. C. Lee and S. A. Zenios, *A fast two-stage algorithm for computing pagerank and it's extension*, SCCM (2003), 15.

[14] D. Cardoso, *Teoria de grafos e aplicações*, Departamento de Matemática, Universidade de Aveiro, 2004/2005.

[15] O. Cardoso, *Recuperação de informação*, Universidade Federal de Lavras (200), 1–6.

[16] O. N. P. Cardoso, *Recuperação de informação*, 2003, pp. 944–945.

[17] O. Frieder D. A. Grossman, *Information retrieval. algorithms and heuristics*, Springer Edition, London, 2003.

[18] S. Velosa D. Pestana, *Introdução à probabilidade e à estatística*, Vol. I, Edições Gulbenkian, Lisboa, 2002.

[19] Jahrbuch Database., *disponível em http://www.emis.de/math/jfm/jmf.html.*

[20] E. K. Donald, *Mathematical typography*, Bull. Amer.Math. Soc (N.S) 1 n.2 (1979), 337–272.

[21] D. Dubin, *Document analysis for visualization*, ACM (1995), 1–6.

[22] S. T. Dumais, *Improving the retrieval of information from external sources*, Behavior Res. Methods Instruments and Computers (1991), 229–236.

[23] L. Martin G. Cleuziou and C. Vrain, *Poboc: an overlapping clustering algorithm. application to rule-based classification and textual data*, in Proceedingsof the 16th Biennial European Conference on Artificial Intelligence (ECAI'04), Valencia, Spain, August (2003), 440–444.

[24] A. Wong G. Salton and Yang, *A vector space model for automatic indexing*, in Communications of the ACM 18(11) (1975), 613–62.

[25] F. Gey, *Models in information retrieval*, SIGIR (1992), Folders of tutorial present at ACM conference.

[26] E. Fernandes Giraldes and V.H Smith, *Curso de algebra linear e geometria analitica*, McGraw-Hill, Portugal, 1995.

[27] F. Guidi, *Searching and retrieving in content-based repositories of formal mathematical knowledge*, Ph.D. Thesis in Computer Science (2003-06), University of Bologna.

[28] F. Guidi and I. Schena, *A query language for a metadata framework about mathematical resources*, The 2nd International Conf. Mathematical Knowledge Management (Feb 2003), Italy.

[29] M. R. Lyu H. Yang, I. King, *Predictive ranking:a novel page ranking aprroach by estimating the web structure*, ACM (2005), 944–945.

[30] D. Harman, *Overview of the third text retriviel conference*, ACM (1993), 1–6.

[31] P. Henderson, *Cadeias de markov*, McGraw-Hill, London, 1993.

[32] P. Calado E. Moura I. Silva, B. Ribeiro-Neto and N. Ziviani, *Linkbased and cotent based evidential inofrmation in a belief network model.*, ACM SIGIR Conference on Research abd Development in Inofrmation Retrieval (2000), 96–103.

[33] R. Motwami L. Page, S. Brin and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical Report 1999-0120, Computer Science Department (1999), Standford University, Standford, CA.

[34] A. N. LangVille and C. D. Meyer, *Deeper inside pagerank*, Internet Math (2003), 229–236.

[35] A. N. Langville and C. D. Meyer, *A survey of eigenvector methods for web information retrieval*, Society for Industrial and Applied Mathematics **47** (2005), 135–161.

[36] D. L. Lee, *Document ranking and the vector-space model.*

[37] R. Lempel and S. Moran, *The stochastic approach for link-struture analysis (salsa) and tkc effect*, ACM (2000), New York.

[38] F. Scarselli M. Bianchini, M. Gori, *Pagerank: A circuital analysis*, in Proceedings of the Eleventh International World Wide Web Conference (2002), 613–62.

[39] F. Mathieu M. Bouklit, *Backrank: an alternative for pagerank?*, ACM (2005), 1122–1123.

[40] M. G. M. Diligenti and M. Maggini, *Web page scoring systems for horizontal and vertical web search*, ACM (2002), 508–516.

[41] MathDi, *disponível em http://www.emis.de/math/di/.*

[42] MathSciNet, *American mathematical society (ams). http://www.ams.org/mathscinet.*

[43] F. MCSherry, *A uniform approach to accelerated pagerank computation*, Microsoft Research, SVC (2001), 575–582.

[44] C. D. Meyer, *Matrix analysis and applied linear algebra*, SIAM Philadelpia (2000), 1–20.

[45] C. D. Meyer and A. N. LangVille, *A reordering for the pagerank problem*, 2003.

[46] B. Miller and A. Youssef, *Technical aspects of the digital library of mathematical functions*, Annals of Mathematics and Artificial Intelligence, v.38 (2003), 121–136.

[47] MoWGLI, *Mathematics on the web: Get it by logics and interfaces. http://mowgli.cs.unibo.it/.*

[48] E. Fredicksen N. Blachman and F. Schneider, *How to do everything with google*, McGraw-Hill, New York, 2003.

[49] B. Ribeiro-Neto E. Moura P. Calado, I. Silva and N. Ziviani, *Linkbased and cotent based evidential inofrmation in a belief network model.*, ACM SIGIR Conference on Research abd Development in Inofrmation Retrieval (2000), 96–103.

[50] G. D. Paulin, *A internet e os seus serviços: Da arpanet à web*, 2007.

[51] V. B. Pinto, *Indexação documentária: uma forma de representação do conhecimento registrado*, Perspect cienc. in. Bela Horizonte, v.6 (2001), 223–234.

[52] L. Pretto, *Link analysis techniques for ranking webpages*, PhD thesis (2002), University of Padua.

[53] B. Ribeiro-Neto and R. Baeza-Yates, *Modern information retrieval*, ACM Press, New York (1999), USA.

[54] S. Robinson, *The ongoing search for efficient web search algorithms*, SIAM News, 37 No. 9 (2004), 4–11.

[55] P. Rodget's, *Thesaurus of english words and phrases*, Longman (1852), London.

[56] R. Motwami S. Brin, L. Page and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, Computer Science Department, Standford University (1999), 353–354.

[57] T. H Haveliwala S. D. Kamvar and G. H. Golub, *Exploiting the block struture of the web for computing pagerank*, Behavior Res. Methods Instruments and Computers (2003), Standford University.

[58] Salton and McHill, *Modelos de recuperação de documentos*, in Proceedings of the Eleventh International World Wide Web Conference (1983), 613–62.

[59] Scopus, *disponível em http://www.scopus.com.*

[60] Mathdex search tool, *disponível em http://www.mathdex.com:8080/mathfind/search.*

[61] M. Sobek, *Google danc: The index update of the google search engine*, Efactory: Internet Agentur KG, http://dance.efactory.d/ (2004), (accessed April 3, 2007.

[62] W. J. Stewart, *Introduction to the numerical solution of markov chains*, Pricenceton University Press (1994), Princeton.

[63] D. Klein-C. Manning T. Haveliwala, S. Kamvar, *Computing pagerank using power extrapolation.*

[64] M. Tastets, *Modelos clásicos de recuperación.*

[65] A. A. Ghorbani W. Xing, *Weighted pagerank algorithm*, CNR'04 (2004), 944–945.

[66] A Youssef, *Roles of maths search in mathematics*, The 15th International Conference on Mathematical Knowledge Management (August 11-12, 2006), Wokingham, UK.

[67] A. Youssef, *Information search and retrieval of mathematical contents: Issues and methods*, The proceedings of the ISCA 14th International Conference on Intelligent and Adaptative Systems and Software Engineering (July 20-22, 2005), Toronto, Canada.

[68] Zentralblatt, *Zentralblatt math database at european mathematical information service (emis). disponível em http://www.emis.de/zmath/.*