Um Modelo Cooperativo e Distribuído para a Recuperação de Informação da WWW

José Exposto[1] António Pina[2] Joaquim Macedo[2]
Albano Alves[1] José Rufino[1]

[1]Instituto Politécnico de Bragança
5301-854 Bragança, Portugal
{exp,albano,rufino}@ipb.pt

[2]Universidade do Minho
4710-057 Braga, Portugal
{pina,macedo}@di.uminho.pt

Resumo

Este artigo apresenta um ambiente da recuperação de informação - SIRe - inteiramente adequado a um espaço multilingue e dinâmico da informação como a Internet. A arquitectura proposta visa assegurar que o sistema pode ser ampliado, para atingir um desempenho mais elevado e melhores resultados das pesquisas, ou reduzido para permitir economizar no sentido de obter uma melhor relação global para o custo/desempenho. Para atingir o almejado desempenho, como uma alternativa de baixo custo às máquinas paralelas tradicionais, o sistema proposto assenta em tecnologias de conveniência para a criação de uma arquitectura de *cluster* baseada em estações de trabalho multi-processadores, ligadas por infra-estruturas de rede de elevado desempenho.

1 Introdução

Hoje em dia, a publicação alargada de diversas fontes de informação através da Internet e os avanços recentes na tecnologia de informação sublinham a importância das bibliotecas electrónicas e de muitos dos temas relacionados. Com a proliferação da informação electrónica, o problema de seleccionar partes de informação relevantes das enormes bases de dados actualmente existentes estão a tornar-se cada vez mais importantes. A recuperação e a filtragem de informação (IR para simplificar) são um domínio de aplicação que representa uma classe, cada vez mais importante, de aplicações comerciais usadas para extrair conhecimento e descobrir as categorias e tendências úteis dos tremendos volumes de dados de informação que estão a ser produzidos pela nossa sociedade de informação.

No final dos anos noventa, a quantidade de informação presente na WWW ultrapassou largamente todas as expectativas, mesmo as mais optimistas. Com a

entrada no novo milénio acentua-se a tendência para um crescimento exponencial. Neste contexto, a pesquisa de informação tem vindo a ocupar um lugar de destaque na rotina de clientes e utilizadores da WWW. Contudo, o aumento exponencial, quer do número de servidores WWW, quer da quantidade de informação disponibilizada em cada um deles, assim como a volatilidade da informação, tem vindo a tornar quase obsoleto o motor de pesquisa tradicional, com uma visão centralizada de recursos: apenas um pequeno número de motores de pesquisa tradicionais continuam a sobreviver, sobretudo graças aos investimentos avultados em equipamentos e tecnologias de suporte computacional e de comunicações.

Por essa razão, a investigação, desenvolvimento e investimento em motores de pesquisa, bem como a concepção de novas arquitecturas computacionais, tem vindo a aumentar consideravelmente nos últimos tempos, por forma a facilitar e melhorar a pesquisa de informação existente, tanto em termos de eficácia como de eficiência. O utilizador, exige, pelo menos, que o motor de pesquisa disponha de uma quantidade grande e abrangente de informação para poder devolver respostas de qualidade e simultaneamente possa manter actualizada aquela informação.

A resposta a estas exigências envolve a criação de um sistema que seja computacionalmente poderoso e capaz em termos de armazenamento e actualização de informação, utilizando recursos e equipamento de investimento reduzido.

Tipicamente, os sistemas de recuperação textuais da informação permitem que os clientes se conectem a uma única base de dados local ou remota. Um sistema IR distribuído deve poder fornecer a um grande número de clientes os acessos simultâneos e eficientes dos múltiplos originais do texto espalhados em locais remotos, especialmente quando cresce o número de clientes e aumenta o número de colecções do texto disponíveis. Para atingir a desejada eficiência são necessárias: 1) arquitecturas de computação e bibliotecas de comunicação adequadas, para extrair desempenho das tecnologias de base; 2) motores de busca eficazes e eficientes e 3) interfaces simples para configurar e administrar o espaço conhecido da pesquisa.

No que se segue apresentamos uma descrição do ambiente SIRe (Scalable Information Retrieval environment) – um sistema paralelo e distribuído, assente em tecnologias de computação baseadas em *clusters*. Na sua concepção está o objectivo de estender o modelo tradicional de Recuperação de Informação, recorrendo à utilização de múltiplas entidades autónomas (SIR) e cooperantes entre si, distribuídas pelos nodos de um ambiente computacional escalável, baseado numa arquitectura de *cluster* que tira partido das tecnologias computacionais e de comunicação de conveniência actualmente existentes.

O modelo proposto tem, também, como objectivo, a definição de uma estrutura lógica de informação que suporta a especificação de aglomerados de conteúdos para permitir o dinamismo de informação e dessa forma garantir a escalabilidade da solução em termos informacionais e de desempenho.

2 Recuperação de Informação tradicional

A Figura 1 apresenta um esquema de Recuperação de Informação tradicional, do qual se destacam dois componentes fundamentais: o robô (Crawler) e o motor de Recuperação de Informação (IRE).

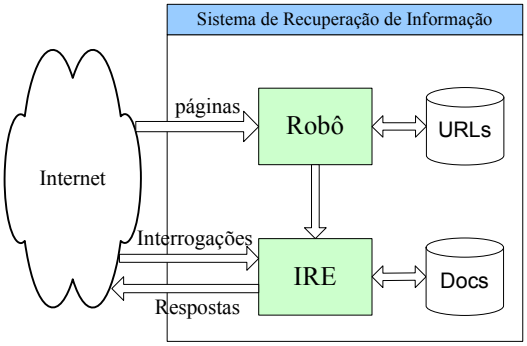

Figura 1: Sistema de Recuperação de Informação tradicional

2.1 Caracterização de um Robô

O robô (Crawler) tem como objectivo descarregar as páginas WWW referenciadas pelos apontadores de moradas de sítios, presentes em cada uma das páginas previamente descarregadas. O algoritmo básico de funcionamento de um robô pode ser descrito através do seguinte encadeamento de acções: 1) retirar um URL de uma lista de moradas disponíveis; 2) resolver o endereço IP do servidor da página referenciada; 3) descarregar a página respectiva; 4) extrair as referências a outras páginas existentes e para cada uma verificar se foi, anteriormente, visitada; 5) caso não tenha sido ainda visitada, é necessário adicionar a referência à lista de URLs ainda a visitar.

Esta visão, simplificada, não deixa, no entanto, de transparecer um conjunto de requisitos indispensáveis ao funcionamento de um Robô num ambiente tão alargado como a WWW.

Eficiência. O algoritmo simplificado do Robô permite, desde logo, descobrir que o acesso à Internet deve ser de banda larga, suficiente para garantir o menor tempo possível no acesso e extracção das páginas; o que vai implicar um desenho para as estruturas de dados que armazenam os respectivos URLs suficientemente eficiente, para suportar a descoberta, a verificação e o armazenamento dos novos URLs em tempo mínimo.

Escalabilidade. O desenho do robô deve ser ajustado a quantidades moderadas de URLs, mas garantir, também, uma resposta eficiente para quantidades

superiores; isto é, o robô deve poder escalar (ampliando e reduzindo) para responder eficientemente ao dinamismo das páginas da web.

Distribuição Estimar o número de páginas da web no espaço de informação alvo, não é uma tarefa fácil. Se consideramos a existência de um total de 10 mil milhões (10.000.000.000) de páginas, a uma média de 8 KB por página, são necessários 74,5 TB de espaço de armazenamento! Segundo um estudo publicado em [8], os motores de pesquisa actuais, cobrem apenas cerca de 30% do valor estimado total para a web; o Google [7] reclama cerca de 3 mil milhões de páginas.

Aquele número pode ser reduzido se considerarmos, apenas, as palavras chave, contidas em cada página, no entanto tal aproximação irá produzir uma sobrecarga substancial nas estruturas de armazenamento e de controlo. De notar que não estamos a entrar em consideração com outros tipos de médias, tais como: imagens, animações flash ou código Java. Mesmo assumindo, apenas, metade daquele valor, cerca de $37,3$ TB, seria necessária um sistema com uma capacidade de armazenamento descomunal, ou então, utilizando sistemas de computação convencionais, com 50 GB de capacidade de armazenamento em disco, seriam necessárias cerca de 763 sistemas daquele tipo.

Dispersão Uma outra questão relevante é a descarga das páginas. Assumindo uma ligação à internet de 5 Mbps seriam necessários quase 4 anos para visitar continuamente os 10 mil milhões de páginas uma só vez. O grau de dinamismo na WWW, implica a descarga de uma mesma página várias vezes durante um determinado tempo de vida. Desta forma, para obter uma imagem instantânea da WWW, 4 anos de visitas é um período de tempo manifestamente incompatível com a realidade.

No caso de usarmos 763 máquinas, cada uma com uma conexão a 5 Mbps, uma visita completa aos 10 mil milhões de páginas poderia ser efectuada em apenas dois dias.

Actualização. De considerável importância é, também, a actualização de páginas que foram já visitadas. O tempo estimado para a visita de todas páginas pode ser tão longo que, durante esse período, se torna obrigatória a re-visita das páginas já, antes, visitadas sob pena da informação presente se tornar obsoleta. Deve, por isso, proceder-se a um escalonamento elaborado de descargas de URLs novos e dos que foram já visitados, pelo menos uma vez.

Delicadeza. A política de delicadeza para com os servidores WWW, embora não sendo um problema estritamente técnico, envolve questões de ética que devem ser rigorosamente consideradas. Devem, também, ser respeitadas as "recomendações" presentes nos ficheiros "robots.txt" dos servidores e nas Meta-etiquetas dos documentos a extrair. Do lado do robô deve, também, ter-se em atenção a taxa de ocupação da ligação à Internet, da instituição hospedeira, podendo haver necessidade de estabelecer limites máximos de ocupação da linha,

ou/e horários de funcionamento do robô para aproveitar, momentos de menor tráfego previsível.

Persistência. Dado o volume dos dados a tratar e o tempo necessário para o respectivo processamento, torna-se necessário um mecanismo de salvaguarda da informação de modo a que, em caso de falha, seja possível uma recuperação graciosa com o mínimo de envolvimento do sistema hospedeiro, evitando-se também o acesso repetido aos servidores WWW para recuperar a informação perdida.

Flexibilidade. Um robô deve poder ser adaptado a vários tipos de aplicações, tais como: recuperação de Informação, estatísticas e *mirroring*.

Extensibilidade. O desenho do robô deve ser extensível de modo a permitir acrescentar novos módulos e/ou configurações que acrescentem novas funcionalidades ou permitam novas especificações; como seria o caso de permitir tratar novos tipos de documentos.

2.2 Motor de Recuperação de Informação

A função do componente motor de Recuperação de Informação (IRE) é devolver ao cliente que faz uma interrogação um lote de localizações de documentos, considerados relevantes, para aquela interrogação.

O IRE recolhe as páginas descarregadas pelo robô e procede à criação dos índices necessários para responder eficientemente às interrogações a que for sujeito. A indexação consiste, basicamente, em analisar as páginas, para poder conhecer a frequência das palavras chave que são armazenadas numa estrutura de dados, designada ficheiro invertido [3].

Este ficheiro é indexado por palavras-chave, contendo cada entrada uma lista dos identificadores das páginas onde a palavra-chave existe, juntamente com a frequência da sua ocorrência na página. Para poder reduzir o número total de entradas no ficheiro invertido são usadas diversas técnicas, tais como, os dicionários negativos e a radicalização; desta forma obtêm-se reduções acumuladas até 70% da quantidade inicial de palavras chave [5].

As interrogações são confrontadas com as representações das páginas, sendo elaborada uma seriação por ordem de relevância, para posterior devolução.

3 Arquitectura SIRe

Quando aplicados à WWW os sistemas de Recuperação de Informação tradicionais apresentam algumas limitações, por razões que se prendem com a dimensão do espaço considerado. O IRE e o Robô são componentes de extrema exigência

de recursos físicos, para os quais existem soluções efectivas, baseadas em arquitecturas centralizadas, que envolvem equipamento extremamente dispendioso.

A evolução dos sistemas de computação e das infra-estruturas de comunicação para a ligação em rede e consequente vulgarização e massificação da sua utilização tem vindo a propiciar a construção de *clusters* baseados em componentes de conveniência, com uma boa relação custo/desempenho. A interligação entre nodos, suportada por tecnologias de alto débito, importadas dos antigos super-computadores, permite ampliar as capacidades dos nodos, tanto em termos computacionais como de armazenamento. Os *clusters*, tipicamente, partilham com os demais sistemas da instituição hospedeira, uma única conexão para a Internet exterior, o que limita, naturalmente, a sua capacidade de ligação à WEB.

Para atingir uma largura de banda aceitável para o acesso à WWW, é de todo conveniente e necessário a dispersão no acesso à Internet. Tal objectivo é alcançável através do desenho de um sistema cooperante que permita orquestrar a actividade de múltiplos *clusters*, individualizados, em que cada um dos quais possui a sua própria ligação à Internet.

Seguindo aquela abordagem, o projecto SIRe, em termos de desenho, assenta na definição de entidades individualizadas, autónomas e dispersas, que cooperam entre si no armazenamento, ou encaminhamento de dados de/e para outras entidades idênticas, com base em decisões administrativas, ou regras de encaminhamento, dinamicamente configuráveis. Cada uma daquelas entidades, designada por SIR, possui capacidades de cooperação múltipla nos domínios da Extracção, da Indexação e do Armazenamento de URLs de páginas WWW, e da resposta a interrogações a clientes sobre a informação textual existente no espaço de informação delimitado. À escala da WWW, para não comprometer a escalabilidade da solução com o aumento, previsível, do número de entidades cooperantes, o sistema global SIRe é organizado numa hierarquia de níveis que associam as entidades individuais (SIR) em agrupamentos lógicos designados por SIRe.

3.1 Requisitos do SIRe

A arquitectura SIRe possui um certo número de propriedades necessárias para dar resposta aos objectivos gerais de um sistema IR de grande escala.

Escalabilidade. Para responder às necessidades actuais de dimensão da WWW e poder adaptar-se, facilmente, quer no que diz respeito à ampliação de capacidade para dar resposta ao crescimento, previsível em numero de páginas, ou em espaço de utilização, quer na capacidade de redução de capacidade para se ajustar a constrangimentos económicos ou diminuição do espaço de utilização alvo.

Dinamismo. Em termos da capacidade de adaptação ao aumento ou redução do número de entidades presentes, durante o tempo de funcionamento do sistema.

Eficiência. Suporte a estruturas de dados robustas distribuídas, para permitir a manipulação de grandes volumes de informação, sem perda de eficiência quando comparada com a manipulação de estruturas de dados centralizadas equivalentes.

Dispersão. A dispersão das entidades do sistema por zonas geográficas distintas, quando alimentada por relações de cooperação apropriadas, pode contribuir para reduzir, significativamente, os inconvenientes das limitações em termos da largura de banda total, acumulada, disponível para o acesso à Internet.

Estruturação. A estruturação é o conceito chave para alcançar a escalabilidade. Por um lado é impensável criar uma estrutura plana de cooperação, entre entidades, por outro lado, a definição de mecanismos de encaminhamento entre entidades, baseados em *hashing*, porque possuem uma localidade mais limitada, diminuem as vantagens da dispersão. Para obviar aqueles inconvenientes podem ser definidas, administrativamente, estruturas hierárquicas de aglomerados lógicos de entidades, com base no conteúdo das páginas, ou na sua localização.

Cooperação. O acesso à WWW é um recurso caro e, por isso, a sua utilização deve ser minimizada, recorrendo sempre que possível ao acesso a recursos internos ao sistema. As entidades que fazem parte do sistema devem cooperar entre si, partilhando, sempre que necessário, a informação de que são responsáveis.

Abrangência. Pretende-se que o sistema cubra a quantidade mais vasta possível de páginas existentes na WWW, não pretendendo impor qualquer limitação de espaço.

Actualização. Face ao dinamismo intrínseco da WWW, o sistema deve permitir uma actualização bem escalonada, sem colocar em causa a descoberta de novas páginas.

Relaxamento. A definição de uma arquitectura distribuída pode dar lugar a tempos de resposta às interrogações superiores aos de um sistema centralizado. Assim, cabe ao cliente a responsabilidade pela definição do grau de cobertura e da qualidade das respostas esperadas, através da definição dos tempos máximos de espera pela resposta às interrogações efectuadas.

Baixo custo. O SIRe apresenta-se como uma alternativa às soluções centralizadas de elevado custo, baseadas em tecnologias e equipamentos proprietários, que tiram partido das tecnologias de conveniência para a construção de um *cluster*, dando o suporte básico às entidades do sistema.

3.2 Entidades do sistema

O sistema SIRe é constituído por três tipos de entidades – a mais elementar (SIR) e duas compostas (SIRe Local e SIRe Alargado) – que se organizam numa hierarquia de níveis que se constituem em topologias. A Figura 2 apresenta as entidades do SIRe e a forma de as associar.

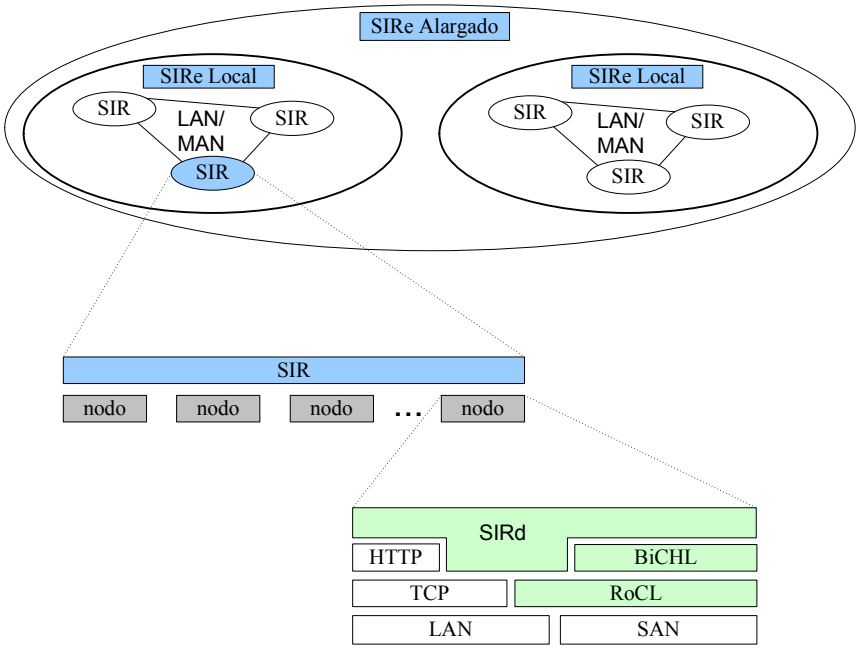

Figura 2: Entidades do SIRe

3.2.1 SIR

O SIR é a entidade básica do sistema que permite a extensão das capacidades computacionais e de armazenamento de um única máquina através da realização como um *cluster*. A utilização de tecnologias de comunicação de elevado desempenho, Gigabit e Myrinet em cada um dos nodos do *cluster*, garantem a eficiência de um conjunto de serviços indispensáveis ao funcionamento do SIR, tais como: o acesso remoto a recursos (RoCL) [1] e o suporte a estruturas de dados distribuídas (BiCHL) [11].

Do ponto de vista externo, o SIR é uma *Single System Image* (SSI), identificável por um único endereço IP. Esta unificação de um agregado de computadores através da referêcnia a um único identificador é possível através da utilização de mecanismos, tais como o *Clone Cluster* [13] ou o *One-IP* [4].

Em termos lógicos um SIR é uma entidade configurável e autónoma. Assim, um SIR dispõe de uma configuração que lhe permite decidir, em cada momento, se um determinado pacote de informação deve ser manipulado localmente ou reencaminhado para uma outra entidade na hierarquia de níveis do sistema.

3.2.2 SIRe Local

Um SIRe Local é definido pela associação de um ou mais SIRs pertencentes à mesma rede institucional (LAN ou MAN), que partilham a mesma linha de acesso externo à Internet. É evidente a falta de dispersão no acesso à Internet exterior no seio desta entidade, uma vez que o acesso é partilhado pelos diferentes constituintes. No entanto, possibilita a uma instituição a criação de um aglomerado de informação mais abrangente e estruturado, podendo o seu conteúdo ser definido de acordo com as necessidades. O SIRe incorpora as definições resultantes da fusão das regras administrativas definidas para cada um dos SIR constituintes. A Figura 3 apresenta um exemplo da interligação necessária para um SIRe Local.

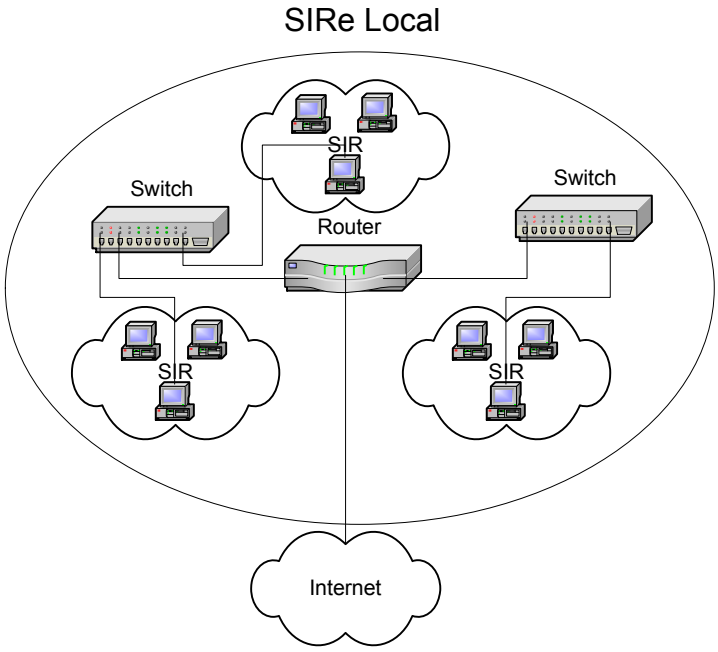

Figura 3: SIRe Local

3.2.3 SIRe Alargado

O SIRe Alargado é constituído pela associação de um ou mais SIRes Locais ou Alargados. O SIRe Alargado para além de permitir a criação de um nível adicional de organização, vem facilitar a dispersão no acesso à Internet. Uma vez que a barreira institucional é quebrada na associação de SIRes Locais, é necessária a definição de políticas de segurança rígidas na constituição de associações entre os diferentes SIRes. A Figura 4 apresenta um exemplo da interligação necessária para um SIRe Amplo.

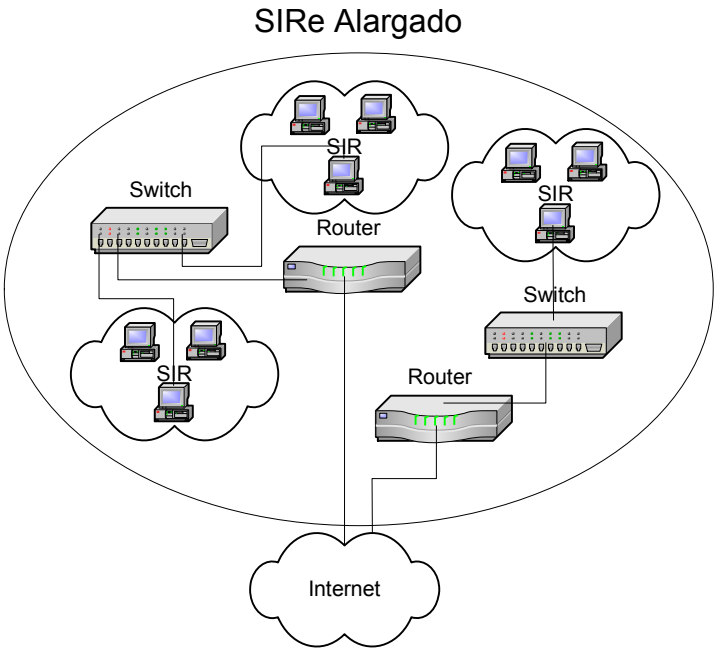

Figura 4: SIRe Amplo

3.3 Instanciação do Sistema

Essencial para o funcionamento integrado de todos os nodos na estrutura SIR é a infra-estrtura de comunicação, de alto débito (SAN), se existente, ou LAN (FastEthernet) usada para suportar o serviço SIRd (ver figura 2). Ainda ao nível da comunicação intra-SIR, um outro serviço essencial é o oferecido pela biblioteca BiCHL que suporte um sistema de páginas de *hash* distribuídas ao nível do SIR para o acesso eficiente e o armazenamento integrado do enorme volume de dados necessários para o funcionamento do SIR. A comunicação inter-SIR é efectuada através de ligações de rede comuns (LAN e WAN). Quando um SIR pretende contactar um outro SIR, este é identificado pelo endereço da sua SSI, que assegura a recepção por apenas um Nodo, escolhido de forma determinista e sem intervenção do emissor.

4 Funcionamento do SIRe

O SIRe distingue-se dos sistemas de Recuperação de Informação tradicionais pela facilidades que dispõe de encaminhamento de informação entre as diferentes entidades constituintes da hierarquia – páginas WEB e respectivas palavras chave, URLs, interrogações e respostas. O SIRe corre em cada um dos SIR constituintes como um serviço distribuído que recorre à identificação da respectiva SSI, para consolidar uma visão unificada e uniforme do SIR (como se este fosse um sistema simples e não um *cluster*).

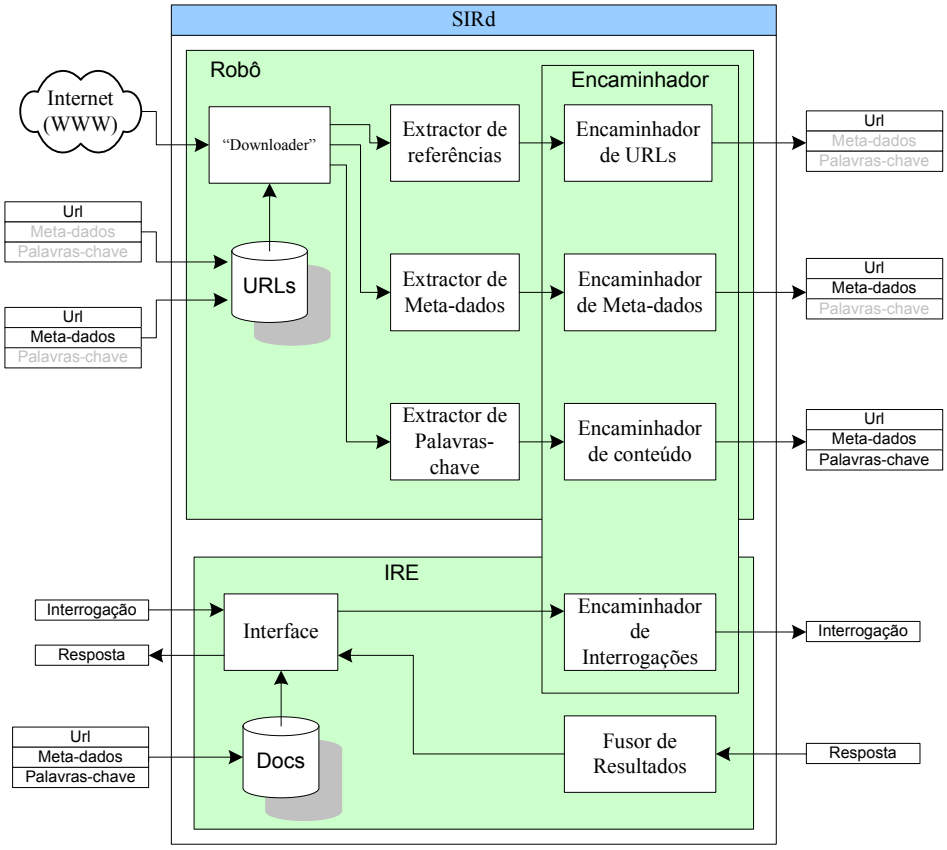

Figura 5: Esquema funcional do serviço SIRd

O esquema funcional do serviço SIRd que corre em cada um dos nodos de um SIR é apresentado na figura 5. Os componentes Robô e o motor de Recuperação de Informação (IRE) são funcionalmente equivalentes aos usados pelos esquemas de Recuperação de Informação tradicionais. O SIRe acrescenta àquele funcionamento a possibilidade da informação colectada pelo robô e das interrogações dirigidas ao IRE serem filtradas por um módulo encaminhador responsável pela, eventual, decisão de encaminhamento da mesma para um novo destino. As caixas dos extremos direito e esquerdo representam o reencaminhamento (saída e entrada) daquela informação para um de três destinos possíveis: 1) o próprio SIR, ou seja, a informação é tratada com base na estrutura de dados distribuída local; 2) um SIR remoto, contactado através do respectivo endereço, ou 3) um SIRe (Local ou Alargado), em que é contactado um dos SIRs seleccionado de entre um dos constituintes desse mesmo SIRe.

No processo de descarga de uma página pelo robô, identificam-se três estados para a informação recolhida: i) o conhecimento do URL, ii) o conhecimento dos meta-dados da página e iii) o conhecimento do seu conteúdo. O conhecimento dos dados de cada estado implica três acções, perfeitamente, separáveis em ins-

tantes diferentes: 1) a extracção das ligações da página descarregada, já previamente encaminhada; 2) a extracção dos meta-dados, o que implica um novo acesso à Internet com, pelo menos, uma operação `HEAD` do `HTTP` [6]; e 3) a extracção das palavras-chave, o que obriga à descarga da página na sua totalidade. As duas últimas acções podem, eventualmente, ser executadas em paralelo recorrendo à operação `GET` do `HTTP`. Esta decomposição em fases de encaminhamento acrescenta um grau de flexibilidade na distribuição da informação, pois conduz à distinção entre as entidades que executam operações tipicamente de descarga e as entidades que executam operações tipicamente de armazenamento de conteúdos, e entidades que, eventualmente, executam ambas as operações.

No que diz respeito ao motor IRE, quando uma interrogação chega a um SIR, esta é difundida, pelo encaminhador de interrogações, para os SIRs remotos com base em resumos de conteúdos disponíveis no próprio encaminhador. Este é um procedimento recursivo que evita a consulta repetida às mesmas entidades, tirando partido da distribuição de carga na operação de interrogação pelas diversas entidades intervenientes. A extensão da distribuição de uma interrogação (SIRs remotos que são consultados) pode aumentar o grau de qualidade e a cobertura das respostas, no entanto, um mecanismo de limitação temporal para a profundidade da pesquisa, é usado para evitar tempos de espera pela resposta, demasiado demorados.

As respostas são fundidas no fusor de resultados e devolvidas ao cliente ou ao SIR que efectuou a interrogação, tomando, obviamente, em consideração a possibilidade de sobreposição de conteúdos [9].

4.1 Tabelas de encaminhamento

A criação e manutenção da topologia de um SIRe, assim como, o encaminhamento de informação, requer a existência de uma tabela de encaminhamento por SIR. Esta tabela armazenada numa estrutura de dados distribuída partilhada por todos os nodos de cada SIR, compreende um identificador (ou endereço físico) e duas secções: a informação topológica e as regras de encaminhamento.

4.1.1 Informação topológica

O SIRe é uma estrutura hierárquica de componentes organizados com base em associações de uma ou mais entidades. Dadas as dimensões em jogo, para manter a escalabilidade do sistema, torna-se necessário que, ao invés de cada entidade dispor de conhecimento total sobre todos os componentes de um dada configuração, ou topologia, se encontrar algum mecanismo que venha a permitir reduzir, substancialmente, os limites desse conhecimento. Assim, cada SIR tem, apenas, conhecimento das entidades ascendentes, na hierarquia de níveis a que pertence e das entidades mais abrangentes e dos descendentes imediatos dos seus ascendentes.

Para exemplificar, se considerássemos uma árvore hierárquica de l níveis, em que cada nível tem c filhos, em que os nodos intermédios são SIRes e os nodos folha são os SIRs, então, iríamos ter c^l SIRs. O que significa que, neste caso,

seria, apenas, necessário o conhecimento de $l \times (c-1)+1$ entidades, para poder conhecer toda a estrutura do SIRe, o que resulta numa ordem de complexidade $O(\log_c(c^l)$. O acesso a um determinado SIR remoto pode implicar uma série de saltos lógicos para atingir o seu destino, o que no pior dos casos corresponde a $\log_c(c^l)$ saltos, ou seja, l, o que é perfeitamente compatível com um número de SIRs igual a c^l, assegurando, desta forma, a escalabilidade do sistema.

Os SIRes enquanto entidades lógicas são, naturalmente, representados por SIRs representantes, de entre os demais constituintes. Se o número de representantes de um SIRe for superior a 1, isso permite a distribuição e balanceamento de carga de encaminhamento, de processamento e, também, alguma capacidade de tolerância a faltas. A utilização de todos os SIRs do SIRe como representantes poderia aumentar, significativamente, a quantidade de informação topológica que cada SIR teria que manter. Assim, considera-se que o ideal é utilizar um número de representantes de ordem logarítmica, obtido através da selecção dos SIRs. A selecção pode fazer-se com base em medidas de qualidade de serviço que garantam a escalabilidade e melhoria de desempenho no acesso às entidades remotas, como é o caso do tempo de resposta das comunicações.

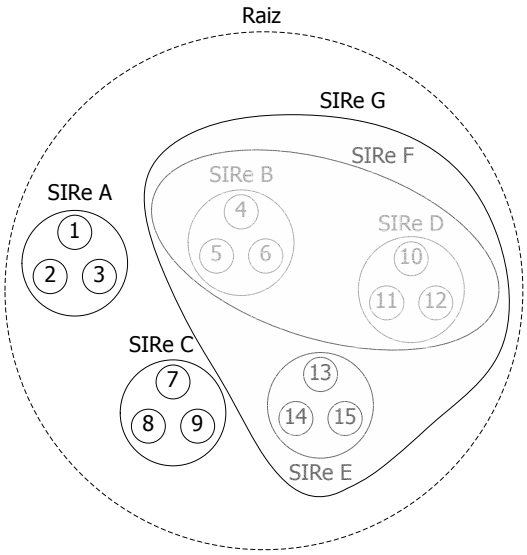

Figura 6: Exemplo de um SIRe

A Figura 6 apresenta um exemplo de uma topologia do SIRe com 11 SIRs (círculos numerados). A, B, C, D e E são SIRes Locais e F e G SIRes Alargados. Em termos de informação topológica, o SIR A teria a seguinte informação topológica:

ID:	SIR 1
Ascendentes:	SIRe A(SIR 1, SIR 2, SIR 3)
	Raiz(SIRe A, SIRe C, SIRe G)
Representantes:	SIRe A: SIR 1
	SIRe C: SIR 7
	SIRe G: SIR 4, SIR 10, SIR 13

O campo de informação dos ascendentes de um SIR reflecte, não só, a enumeração das entidades visíveis mas, também, o respectivo, encadeamento estrutural.

4.1.2 Regras de encaminhamento

No SIRe, as regras de encaminhamento são o mecanismo usado para suportar a estruturação do sistema e a possibilidade de cooperação entre as entidades constituintes. Em cada SIR, a definição das regras de encaminhamento e a consequente afectação aos SIRes compostos, por disjunção lógica das regras dos SIRs, responsabiliza aquela entidade pelo processamento da informação.

No robô, com base no estado de informação disponível, distinguem-se três níveis de regras: 1) as regras de URLs, 2) as regras de meta-dados e 3) as regras de conteúdo. As primeiras são de definição obrigatória, uma vez que existe a necessidade de atribuir os URLs a, pelo menos, um dos SIRs, para que seja efectuada a descarga da página correspondente. As restantes regras são opcionais e, em caso de omissão, o conteúdo da página é colocado no SIR a que foi originalmente destinado o URL.

No exemplo da Figura 6 se supusermos que para cada SIR n está definida a expressão R_n para as respectivas regras, o SIR 1 teria as seguintes regras de encaminhamento:

R_1	SIR 1
R_2	SIR 2
R_3	SIR 3
$R_7 \vee R_8 \vee R_9$	SIRe C
$R_4 \vee R_5 \vee R_6 \vee R_{10} \vee R_{11} \vee$	
$R_{12} \vee R_{13} \vee R_{14} \vee R_{15}$	SIRe G

A definição de cada uma das regras é baseada numa linguagem simples de predicados que usa, para criar a expressão (fórmula) da regra, os operadores lógicos: + ($\vee$), * ($\wedge$) e ! ($\neg$). Cada predicado, previamente definido, fornece o suporte para as operações básicas de manipulação: de números, de datas, de cadeias de caracteres e de URLs.

A título de exemplo, o predicado `domain(X, Y)` verifica se Y é o domínio do URL X; o predicado `expires(X, Y)` verifica se Y é a data de expiração de uma página com o URL X; e o predicado `lt(X, Y)` verifica se X é menor do que Y. Com base naquela informação, um administrador pode definir a seguinte

regra de encaminhamento para um SIR: `domain(URL, "pt") * expires(URL, Y) * lt(Y, "26/05/2003 23:59")`, sendo interpretada como aceitar os URLs do domínio `pt`, com data de expiração anterior a `"26/05/2003 23:59"`.

Embora este mecanismo não seja compatível com a definição dos predicados feita pelo administrador, garante que a predefinição de um conjunto de predicados que suportem as operações básicas referidas, é suficientemente abrangente para o tipo de regras previsto. No entanto, o administrador dispõe da flexibilidade necessária para poder definir as expressões que desejar, combinando os predicados com os operadores lógicos, de modo a especificar as particularidades de cada SIR.

4.2 Construção da topologia

A criação de novas entidades é um processo iterativo, de adição sucessiva de entidades. Em primeiro lugar, a construção de uma topologia envolve a definição das estruturas físicas, ou seja, a definição e preparação dos nodos de um *cluster* e a instalação dos serviços, anteriormente, apresentados.

A definição de um SIR consiste na atribuição de uma identificação e na especificação das regras que indicam qual a informação a ser processada por aquele SIR.

A definição das entidades compostas (SIRes) implica: i) a atribuição de uma identificação à nova entidade, ii) o envio do pedido de composição e iii) a recepção do pedido de composição pela entidade hospedeira. Para efeitos de actualização da informação topológica mantida pelas entidades no sistema, são usadas *spanning trees* para difundir a nova definição. A sobrecarga de comunicações associadas à difusão pode ser atenuada se for actualizada, apenas, a informação mantida pelos representantes das entidades compostas.

Voltando ao exemplo da Figura 6, consideremos a criação dos SIRs 13, 14 e 15 e a sua composição no SIRe E, seguida da composição do SIRe E e F para a criação do SIRe G. Assim, são preparados 3 *clusters*, em que se instalam os serviços necessários e atribuídos os identificadores e as regras prórias. Seguidamente, por composição, é criado o SIRe E (Figura 7).

Para criar o SIRe G, um dos SIRs no SIRe E (por exemplo, o 13) contacta um SIR no SIRe F (por exemplo, o 4). O SIR 4 responde ao 13 com a sua visão da topologia enviando, ao mesmo tempo, os pedidos de actualização da nova entrada, aos representantes das entidades de que tem conhecimento.

A remoção de entidades de um SIRe segue a mesma filosofia de actualização. Após a remoção, a capacidade do sistema no seu todo fica reduzida. No entanto, não é prejudicada a sua consistência e funcionamento. A entidade retirada mantém a sua autonomia podendo continuar a funcionar isoladamente e, eventualmente, vir de novo a associar-se ao anterior sistema ou a outro existente numa outra configuração.

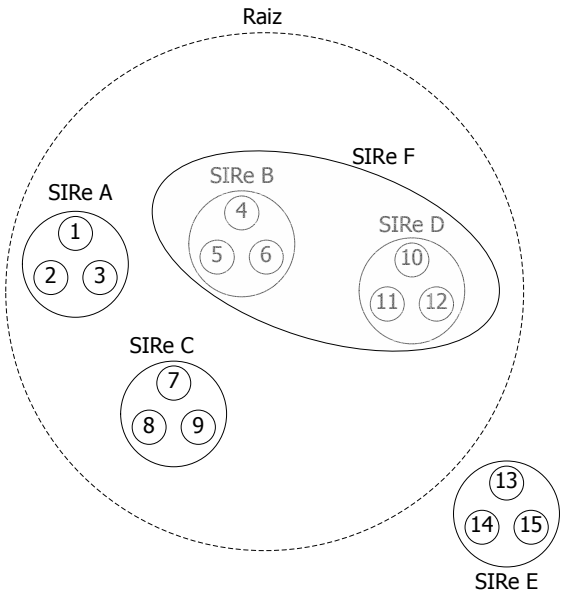

Figura 7: Composição de SIRes

5 Conclusões

O projecto SIRe tem como origem a necessidade de responder de forma apropriada às cada vez maiores exigências que se colocam à pesquisa de informação na era da Internet e da globalização. Dentro das suas características e objectivos sobressai a necessidade de dar uma resposta à crescente demanda, eficaz em termos de qualidade dos resultados devolvidos e eficiente em termos da relação custo/desempenho. A abordagem passa pela construção de um sistema escalável paralelo e distribuído que recorre a componentes de conveniência interligados por tecnologias de comunicação de alto débito que garantem a realização das infra-estruturas que suportam as tecnologias de base e os paradigmas de computação e comunicação mais apropriados. Um outro desígnio essencial é a escalabilidade, entendida no sentido da adaptação ao crescimento ou à redução do volume de dados a tratar, de acordo com a minimização da relação abrangência/custo.

A capacidade organizativa em aglomerados de conteúdo confere, a cada instituição, a possibilidade de decidir sobre a informação a manipular, de forma a rentabilizar recursos de que dispõe (equipamento, acesso à WWW, etc.). Desta forma, através da criação de políticas e mecanismos de cooperação inter-institucional, para a partilha e replicação de conteúdos, assentes numa estrutura hierárquica podem, vir a garantir-se a minimização dos custos de acesso à WWW, tanto em termos económicos como em termos do tempo total despendido nas tarefas gerais de armazenamento e extracção de informação.

A utilização de estruturas de dados distribuídas e serviços de suporte à manipulação de recursos, assentes em tecnologias de *cluster*, são altamente valiosas, pois permitem a sua unificação e abstracção e, desta forma, estendendo a capaci-

dade de cada nodo individualizado em termos computacionais/comunicacionais e de armazenamento.

A dispersão geográfica dos SIR é, simultaneamente, vantajosa no aproveitamento do acesso disperso à WWW e na possibilidade de distribuição e balanceamento de carga computacional e de armazenamento, sem esquecer o aumento da capacidade de tolerância a faltas.

O mecanismo encontrado para a representação topológica oferece uma escalibilidade compatível com as necessidades de crescimento do número de entidades presentes. Os predicados disponibilizados para a definição das regras de encaminhamento, embora limitados na sua diversificação, possuem uma cobertura e uma flexibilidade que podemos considerar suficientes para o tipo de regras previsto.

O desenho do robô é de extrema importância para o sucesso do SIRe. O algoritmo a utilizar, nomeadamente, no que diz respeito ao escalonamento da visita a novas páginas e a revisita a páginas previamente visitadas, deve ser cuidadosamente elaborado, de modo a permitir a descarga eficiente das páginas, tal como apontam os trabalhos referidos em [10, 12, 2].

O encaminhamento de informação e a definição de SIRes Alargados que utilizam conexões WAN pode considerar-se um ponto crítico devido ao tipo e instabilidade, intrínseca, das conexões estabelecidas através da Internet. A utilização de métricas de qualidade de serviço e o cálculo ou estimativa da proximidade entre as entidades é por, essa razão, um assunto alvo de uma investigação cuidada e de propostas arrojadas. Finalmente, a fusão de resultados deixa antever algumas dificuldades, devido à existência de documentos duplicados em diferentes entidades. O trabalho de investigação citado em [9] está a ser usado como ponto de referência para a descoberta de novos caminhos para resolução deste problema.

Referências

[1] A. Alves, A. Pina, J. Rufino, and J. Exposto. RoCL: a resource oriented communication library. In *International Conference on Parallel and Distributed Computing (Euro-Par '03)*, 2003.

[2] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proc. of the 11th International World–Wide Web Conference*, 2002.

[3] C.J. van Rijsbergen. Information Retrieval, 1979.

[4] O. P. Damani, P. E. Chung, Y. Huang, C. Kintala, and Y.-M. Wang. ONE-IP: Techniques for hosting a service on a cluster of machines.

[5] J. Exposto. Aglomeração não Hierárquica em Sistemas Distribuídos de Recuperação de Informação. Master's thesis, Universidade do Minho, Dezembro 1997.

[6] R. Fielding, J. Gettys, J. C. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. Hypertext transfer protocol – http/1.1. Request for Comments 2616, June 1999.

[7] Google. http://www.google.com.

[8] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.

[9] J. M. H. Macedo. *Recuperação de Informação Textual Distribuída por Fontes Múltiplas com Sobreposição*. PhD thesis, Universidade do Minho, 2001.

[10] M. Najork and A. Heydon. On high-performance web crawling, 2001.

[11] J. Rufino, A. Pina, A. Alves, and J. Exposto. Distributed Paged Hash Tables. In *5th International Meeting on High Performance Computing for Computational Science (VECPAR '02)*, 2002.

[12] V. Shkapenyuk and T. Suel. Design and implementation of a high-performance distributed web crawler. In *ICDE*, 2002.

[13] S. Vaidya and K. J. Christensen. A single system image server cluster using duplicated mac and ip addresses. In *IEEE 26th Conference on Local Computer Networks*, pages 206–214, 2001.