Análise da Disponibilização de um Índice Invertido em P2P

Nuno Lopes, Carlos Baquero

Grupo de Sistemas Distribuídos
Departamento de Informática, Universidade do Minho

Resumo

A procura de informação com base em conteúdos é uma funcionalidade fundamental na construção de sistemas de partilha de ficheiros e recursos. Contudo, as soluções eficientes para a partilha *peer-to-peer* que têm por base DHTs - *distributed hash tables* - não são capazes de efectuar procuras devido ao seu modelo de funcionamento com base em identificadores unívocos.
Este trabalho apresenta um modelo de construção de um índice invertido que permite a execução de buscas eficientes nos sistemas DHTs.

Os Sistemas Peer–2–Peer

Os sistemas P2P permitem a partilha de informação através de mecanismos autónomos e descentralizados.
A primeira geração destes sistemas partilhava ficheiros que eram encontrados através da procura de palavras. A rede Gnutella é um exemplo destes sistemas, na qual o processo de procura assenta em técnicas de disseminação epidémica, o que provoca uma sobrecarga nos recursos de comunicação.
A segunda geração de sistemas P2P é baseada no conceito de DHT, que faz um uso eficiente do meio de comunicação, mas não é capaz de efectuar procuras.

Porquê um Índice Invertido?

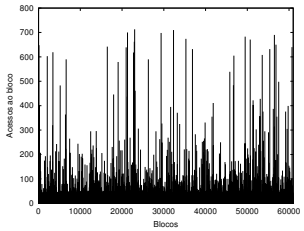
Como os novos sistemas não são capazes de efectuar buscas, é necessário permitir a procura de documentos sem afectar as propriedades de escala e eficiência destes sistemas. Um índice invertido distribuído é utilizado para obter as referências dos documentos onde uma palavra específica ocorre. Mas porque a ocorrência de palavras nos documentos segue uma distribuição Zipf, *um índice invertido não pode ser aplicado directamente sobre um DHT*, uma vez que tal conduz a sobrecargas nos nodos.

Abordagem

Um índice invertido pode ser decomposto numa associação de palavras a conjuntos de referências para documentos. Para implementar este modelo foram utilizadas árvores-B+ sobre um sistema DHT.
O papel do DHT é oferecer um suporte base para blocos de informação com tamanho constante. As árvores-B+ são responsáveis pelo agrupamento desses blocos de modo a permitirem uma implementação escalável e adaptativa de conjuntos de referências.

Resultados (1)

- O número de acessos verificado por bloco é bastante assimétrico e relaciona-se com a frequência das palavras. Alguns blocos são alvo de contenção porque contêm a raiz das árvores com as palavras mais frequentes.



Resultados (2)

- A utilização de *caching* em cada nodo reduz em uma ordem de grandeza os acessos aos blocos mais requisitados e uniformiza os restantes acessos.