

# 1 Introdução

No passado, uma das maiores dificuldades na hora da realização de algum projeto ou trabalho era encontrar subsídios de conteúdo para o seu desenvolvimento. A fase de pesquisa, de descoberta de bibliografia e de coleta de trabalhos relacionados consumia um tempo considerável. Porém, tal fato vem mudando cada vez mais a cada dia. A falta de informação vem sendo rapidamente substituída por excesso de informação: basta entrar na Internet e realizar alguma consulta sobre um tema qualquer numa máquina de busca para que se tenha uma boa idéia não só do cenário atual de riqueza de informações, como também do cenário futuro.

A facilidade de publicação de informações na web é, sem dúvida, o principal fator responsável pelo rápido avanço da ciência nos dias de hoje. A troca de informações entre pessoas se dá de forma muito dinâmica, possibilitando trabalhos em conjunto e facilitando a cooperação. A Figura 1 ilustra a rápida taxa de crescimento de páginas (documentos HTML) vivida pela Internet nos últimos tempos.

É muito comum o acúmulo de diversos documentos durante a realização de projetos ou trabalhos em equipe. Porém, o conhecimento adquirido por um membro de uma equipe dificilmente é transmitido para outros membros da mesma equipe, pior ainda se considerarmos a transmissão entre equipes de diferentes projetos. Isso porque o gerenciamento de tais documentos não é bem feito – poucas são as organizações que possuem algum tipo de mecanismo de gestão de conhecimento (knowledge management) [1].

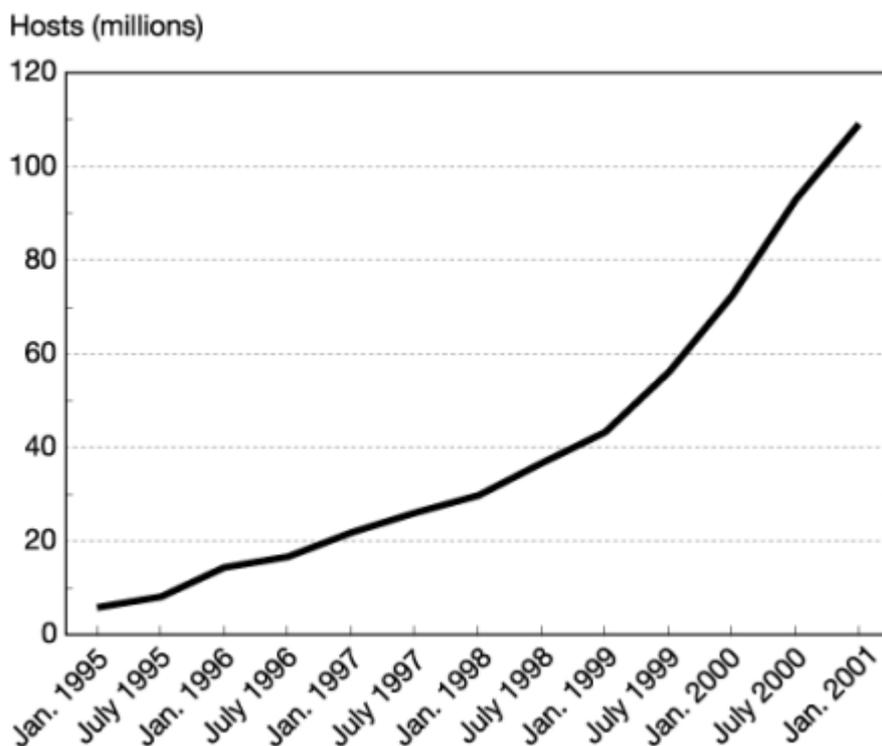


Figura 1: Evolução do número de hosts na Internet até Janeiro de 2001

Ocorre, então, que o conhecimento adquirido ao longo do tempo não é da organização, mas sim das pessoas que a compõem. Se algum dia alguém deixa a organização, esta perde todo o conhecimento representado pela pessoa que está partindo – todo esforço que possa ter sido envolvido (investimento em treinamento, especialização, tempo de integração com a empresa, etc) é completamente perdido. É interessante que haja formas de uma organização manter o conhecimento mesmo quando houver rotatividade de seus integrantes – esse vem sendo um campo que vem recebendo investimentos maciços [1].

É no contexto de auxiliar a descoberta e o gerenciamento do conhecimento acumulado por equipes que nossa ferramenta se encaixa. Esse trabalho tem por objetivo propor um *framework* para um sistema multi-agentes pró-ativo para busca e classificação de documentos textuais. Entende-se por documento textual um documento cujas principais informações estejam descritas no formato texto – por exemplo, não é do escopo desse trabalho trabalhar informações residentes em documentos multimídia, que possuem áudio ou vídeo, ou mesmo em documentos cuja interpretação das figuras é decisiva para a sua compreensão.

O objetivo é deixar independentes as questões de plataforma (busca e gerenciamento de documentos) e do algoritmo de classificação utilizado, dando ao instanciador do *framework* a possibilidade de levar em conta características específicas do domínio dos documentos categorizados na seleção do algoritmo a ser utilizado. Por exemplo, é possível utilizar algoritmos já existentes, porém com adaptações (heurísticas) segundo o domínio, ou mesmo o desenvolvimento de algoritmos inteiramente novos.

Um claro exemplo de como a característica de flexibilidade em relação a algoritmos selecionados, aliada a tipos de documentos processados, pode ser utilizada é num ambiente em que seja possível ao autor adicionar informações sobre a natureza do documento gerado (meta-informação) – por exemplo, adicionar informações referentes à sua categoria, data de criação, data de última alteração, organização que o gerou, além do nome do próprio autor, somente para citar algumas. Tais informações poderiam (e deveriam) ser utilizadas, em conjunto com o conteúdo do documento, para determinar a melhor classificação de acordo com um grupo de categorias. Um ambiente em que tais condições são satisfeitas é o da *web semântica* (*semantic web*) [2].

O restante do documento está estruturado da seguinte forma: o capítulo 2 fala um pouco sobre a área de gestão de conhecimento, o capítulo 3 fala da *web semântica*, que é uma reengenharia da *web* existente hoje que vai trazer muitos ganhos para a área de classificação de documentos, o capítulo 4 trata de agentes de software, o capítulo 5 fala sobre algoritmos de geração de categorias (que poderiam ser utilizados para alimentar um algoritmo de classificação baseado em categorias). Um leitor que já tenha experiência com essas tecnologias poderia se dirigir diretamente aos capítulos de 6 a 9, que são os principais, onde é descrito todo o trabalho realizado: no capítulo 6 há uma visão geral do *framework* desenvolvido, o capítulo 7 descreve o sistema multi-agentes do *framework*, o capítulo 8 é um manual de instanciação e o capítulo 9 descreve as instanciações realizadas. O capítulo 10 descreve alguns trabalhos relacionados, e concluindo, o capítulo 11 apresenta algumas conclusões e possíveis trabalhos futuros.