Uma Arquitetura de Agentes Cooperativos de Informação para a Web Baseada em Ontologias

Fred Freitas[1], Tércio de Morais Sampaio[1], Rafael Cobra Teske[2], Guilherme Bittencourt[2]

[1]Departamento de Tecnologia de Informação – Universidade Federal de Alagoas
Campus A.C. Simões - BR 104 - Km 14 - Tabuleiro dos Martins – 57.072-970
Maceió – AL – Brasil

[2]Departamento de Automação e Sistemas - Universidade Federal de Santa Catarina
Caixa Postal 476 - 88.040-900 - Florianópolis - SC – Brasil

fred.freitas@tci.ufal.br,terciomorais@yahoo.com,{cobra,gb}@das.ufsc.br

Abstract. *In the Web, there are classes of pages (e.g., call for papers' and researchers' page), which are interrelated forming clusters (Science). We propose a reusable architecture for multi-agent systems to retrieve and classify pages from these clusters, supported by data extraction. Crucial requirements: (a) a Web vision coupling this vision for contents to a functional vision (role of pages in data presentation); (b) ontologies to represent agents' knowledge about tasks and the cluster. This web vision and agents' cooperation accelerate retrieval. We got quite promising results with two agents for the page classes of scientific events and articles. A comparison with WebKB comes up with a new requirement: a detailed ontology cluster.*

Resumo. *Existem, na Web, classes de páginas (e.g. "call for papers", pesquisadores) que inter-relacionam-se, formando* grupos *(o meio científico). Propomos uma arquitetura reusável de sistemas multiagentes cognitivos para recuperar e classificar páginas destes grupos, baseada na extração de dados. Requisitos: (a) uma visão da Web que acopla a visão por conteúdo a uma visão funcional (papel das páginas na apresentação de dados); (b) ontologias sobre as tarefas dos agentes e o* grupo. *Esta visão da Web e a cooperação entre agentes aceleram a recuperação. Obtivemos bons resultados com dois agentes para as classes de eventos e artigos científicos. Uma comparação com o WebKB sugere um novo requisito: uma ontologia detalhada do* grupo.

1. Introdução

Apesar do termo "agentes cooperativos de informação" ser muito citado, há poucos sistemas na Web que mostram alguma forma de cooperação ou integração entre tarefas relacionadas a texto, como recuperação, classificação e extração de dados. Na definição de manipulação integrada de informação, a extração é mencionada como técnica de aquisição de conhecimento, sugerindo implicitamente integração entre as tarefas. Os sistemas de extração atuam sobre domínios muito restritos, como notícias sobre terrorismo, classificados. Um fato negligenciado sobre as classes de páginas processadas por extratores é que muitas delas inter-relacionam-se, formando *grupos* (*clusters*).

Visando integração e cooperação, projetamos uma arquitetura de sistemas multi-agentes que recupera e classifica páginas de grupos de classes inter-relacionadas na Web, extraindo dados delas. Para permitir a cooperação, dois requisitos são cruciais: (a) uma visão da Web que acopla a visão por conteúdo a uma visão funcional (papel das páginas na apresentação de dados); (b) ontologias sobre as tarefas dos agentes e o *grupo*. São relatados promissores resultados com os agentes de eventos e artigos científicos, com. A categorização funcional, e as listas em particular, melhoram a busca e uma comparação com o WebKB sugere o uso de uma ontologia detalhada do *grupo.*

O artigo está assim organizado: A Seção 2 descreve a visão da Web proposta. A seção 3 introduz a arquitetura e seus componentes. A seção 4 apresenta um estudo de caso, o sistema MASTER-Web (*Multi-Agent System for Text Extraction, classification and Retrieval over the Web*), aplicado ao grupo científico, com dois agentes, um de eventos e outro de artigos científicos. Os resultados nas classificações por conteúdo e funcional são apresentados. A seção 5 compara o MASTER-Web com sistemas similares, e a seção 6 traz trabalhos futuros e conclusões.

2. Visão da Web para a Manipulação Integrada de Informação

As páginas que apresentam entidades compartilham estilo de editoração, , terminologia e o conjunto de atributos. A criação de extratores baseia-se neste fato, na existência de *classes de páginas*, definindo uma visão da Web por conteúdo. Em páginas de pesquisadores, por exemplo, encontram-se dados como instituições, áreas de interesse, artigos e muitos outros itens. O conjunto de atributos de uma classe é *discriminante,* no sentido em que sua presença ajuda a distinguir instâncias da classe [Rilloff 94].

Muitos ponteiros das páginas que pertencem a estas classes apontam para outras páginas contendo entidades ou atributos ou âncoras pertencentes a um número reduzido de outras classes. Chamamos a este conjunto de classes *grupo (cluster) de classe*. Por exemplo, em páginas de pesquisadores, com certeza serão encontrados ponteiros para páginas de artigos, podem ser localizados ponteiros para chamadas de trabalho de eventos científicos, e páginas de outras classes.

Uma visão alternativa da Web diz respeito à funcionalidade das páginas, dividindo-as de acordo com o seu papel na ligação e na apresentação dos dados. Visando a extração integrada, as categorias funcionais se dividem em: *páginas-conteúdo,* (membros de uma classe), *listas* de páginas-conteúdo, *mensagens* (sobre uma classe), *recomendações* (páginas de outras classes), ou *lixo*.

Tanto para identificar precisamente as páginas com instâncias das classes processadas, como para localizá-las rapidamente, beneficiando-se dos relacionamentos entre elas refletidos nas âncoras, as duas visões devem ser usadas simultaneamente.

3. Arquitetura Proposta

Propomos uma arquitetura de Sistemas Multiagentes Cognitivos [Freitas & Bittencourt 2003] para resolver o problema da extração integrada de páginas-conteúdo pertencentes às classes que integram um grupo (*cluster).* A motivação principal para o emprego de sistemas multiagentes é beneficiar-se dos relacionamentos entre as classes. A visão geral da arquitetura está ilustrada na figura 1.

Cada agente, representado como um círculo na figura, reconhece, filtra e classifica páginas que, supostamente, pertençam à classe de páginas processada por eles (por exemplo, páginas de CFPs, artigos e outros para o grupo científico), extraindo também seus atributos. Uma vez que os agentes cooperam, possuindo, porém, responsabilidades distintas, a arquitetura baseia-se na abordagem de Resolução Distribuída de Problemas (RDP) [Álvares & Sichman 95]. A estrela indica troca de mensagens contendo regras de reconhecimento e fatos (conhecimento dos agentes), além das URLs sugeridas entre os agentes.

Cada agente possui um meta-robô, que se conecta a múltiplos mecanismos de busca - como *Altavista, Excite, Infoseek* e outros. Ele consulta os mecanismos de busca com palavras-chave que garantem cobertura em relação à classe de páginas processada pelo agente. (e.g., os termos *'call for papers'* e *'call for participation'* para o agente CFP). Devido à falta de precisão, o conjunto de páginas resultante das consultas recai em vários grupos funcionais além do de páginas-conteúdo tratado, apresentando muitas listas, mensagens, páginas-conteúdo de outras classes, e lixo. As URLs são dispostas numa fila de URLs de baixa prioridade. Cada agente continuamente acessará, além desta fila, outra de alta prioridade, que armazena URLs sugeridas por outros agentes ou presentes em páginas da categoria funcional listas. Afinal, estes endereços são obtidos dentro de um contexto mais confiável e com maior probabilidade de ser relevante do que as listas de resultados dos mecanismos de buscas.

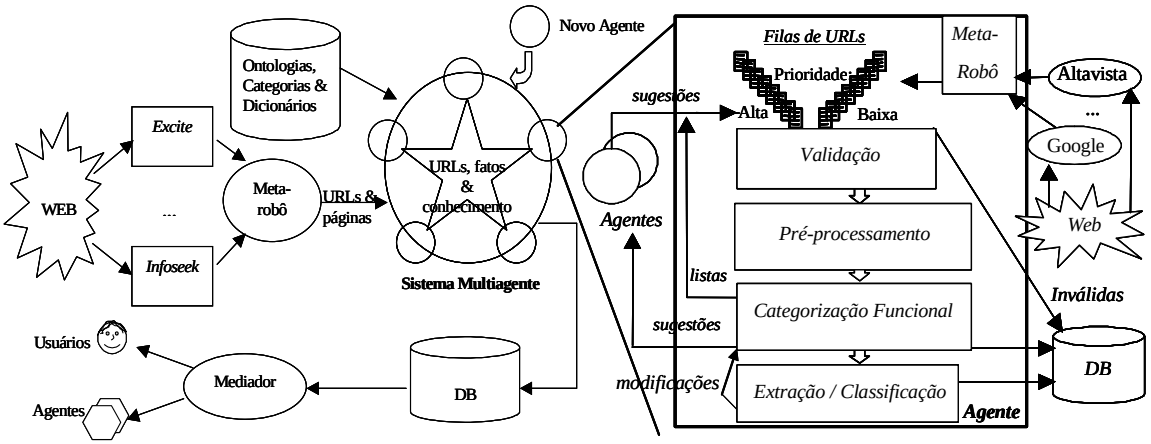

Figura 1. Visão geral da arquitetura de sistemas multiagentes para manipulação integrada, mostrando o funcionamento de um agente em detalhe.

Um mediador estará disponível, com a função de ajudar às consultas aos dados, provendo visões não-normalizadas - mais simples - da base de dados, e permitindo a qualquer usuário ou agente beneficiar-se do acesso aos dados extraídos.

Ao entrar no sistema, os agentes registram-se e anunciam-se aos outros agentes, mandando fatos e regras de reconhecimento de páginas e ponteiros úteis a si próprio, que serão empregadas pelos outros para lhe indicarem sugestões de páginas. O novo agente receberá, reciprocamente, regras úteis aos outros agentes. Assim, quando um agente acha informação que dispara alguma das regras referentes aos outros, este agente repassa a informação (ponteiro ou página) ao agente que lhe enviou a regra disparada.

3.1. Tarefas dos Agentes

Um agente executa quatro tarefas em cada página que processa:

- Validação: Nesta fase são eliminadas páginas inacessíveis, já existentes no BD e em formatos que os agentes não possam processar.

Pré-processamento: Representa as páginas de várias maneiras, tais como conteúdo com e sem HTML, palavras-chave e freqüências, ponteiros, e-mails, e outros, com dados extraídos delas, aplicando, se necessário, recuperação de informação e processamento de linguagem natural (PLN). Os dados passam ao motor de inferência.

- Categorização Funcional: Aqui, as páginas são classificadas em grupos funcionais e são encontradas e enviadas as sugestões para outros agentes, quando uma das regras enviadas por eles dispara. Por exemplo, uma âncora com a palavra "*conference*" é útil para o agente CFP. As sugestões podem, inclusive, acionar buscas em diretórios com prefixo comum, como `/staff/` para o agente de pesquisadores.

- Extração e Classificação: São extraídos os atributos, armazenados na base de dados, ou, pela inconsistência ou inexistência destes, corrigida a classificação de página em relação aos grupos funcionais. Por exemplo, a presença de datas com mais de um ano de intervalo numa página de CFP, denunciam uma página confundida por um CFP.

Lançando mão das representações necessárias, a extração é efetuada por uma combinação de *templates*, e regras ou uma categoria é inferida, normalmente quando achados termos dos dicionários (e.g. a presença de siglas de estados norte-americanos). Após a extração, os atributos podem, adicionalmente, ser formatados (e.g., datas). A partir daí, testam-se casos que contém um conjunto mínimo de atributos para identificar páginas conteúdo. Um exemplo está disposto na próxima seção.

Após a identificação de uma página conteúdo, outros dados são procurados e extraídos. Até este ponto, os dados extraídos apenas evidenciavam a existência de determinados itens de informação. Entretanto, os dados extraídos não são devidamente contextualizados. Por exemplo, datas extraídas de páginas de CFP podem ter significados diversos, como data limite para entrega de trabalhos – *deadline*, data de notificação, data do evento, etc.

Na etapa de extração, combinações de diversos dados extraídos podem compor informações cuja extração é desejada. Para isto, são usadas instâncias de *templates* onde são definidos quais dados se relacionam e como ocorre esta relação de modo a formar uma informação consistente. Um exemplo de extração encontra-se na seção 4.

3.2. Conhecimento dos Agentes

Ontologias desempenham um papel fundamental na arquitetura, servindo não só como vocabulário de comunicação entre agentes, como também na definição e organização apropriadas de conceitos, relações, e restrições. Quatro ontologias se fazem necessárias:

- Ontologia do domínio (ou grupo): Ontologia principal, devendo ser bastante detalhada para garantir precisão à classificação por conteúdo (ver subseção 5.1.).

- Ontologia da Web: Contém definições de *hyperlink*, termo e freqüência, e de página da Web em suas várias representações e atributos - como listas de palavras-chaves e

freqüências, ponteiros, e-mails, etc. Pode, ainda, conter definições relativas à Internet, como protocolos e tipos de arquivos, além de representações de páginas em PLN.

- Ontologia de manipulação integrada de informação: Classes e instâncias empregadas na extração e classificação funcional e por conteúdo. Inclui

– *Templates* reconhecedores das categorias funcionais e páginas-conteúdo

– *Templates* extratores e classificadores de dados

– Classes auxiliares, como meta-definições de conceitos e sinônimos e palavras-chave, classes de PLN (tendo como atributos *parts-of-speech-tags como* rótulos de frases/sintagmas), agentes e habilidades, etc

– Casos complexos que identificam atributos, classes de páginas, categorias funcionais e sugestões para outros agentes. Os casos devem ser bastante expressivos, com conjuntos de atributos e conceitos cuja presença e/ou ausência implica que em categorização como página-conteúdo. Segue abaixo o exemplo do caso mais comum em chamadas para eventos científicos ao vivo (conferências e *workshops)*, em que uma página apresenta no seu início os atributos data inicial do evento e localização (país do evento) e algum termo relacionado ao conceito de evento ao vivo, como as expressões *"call for papers" (*por herança*), "conference"* ou *"workshop"*. Uma regra associada aos atributos do caso (e.g. `Slots-in-the-Beginning)` é disparada se as condições são atendidas.

```
([Date-time-in-the- beginning] of Case
      (Slots-in-the-Beginning [Initial-Date] [takes-Place-at])
      (Concepts-in-the-Beginning [live-scientific-event]))
```

- Ontologias auxiliares: Conhecimento útil de outras áreas de conhecimento. Ontologias lingüísticas, como o WordNet [Miller 95], de tempo e locais, além de outras específicas de um agente (como dados bibliográficos para o agente de artigos científicos).

4. Estudo de caso: o grupo científico

A ontologia do domínio científico [Freitas 2001] foi reusada a partir da ontologia do projeto europeu $(KA)^2$ (*Knowledge Annotation Initiative of the Knowledge Acquisition Community*) [Benjamins et al 98], refinada em vários aspectos. O principal deles foi a inclusão de classes abstratas – que não contêm instâncias –, visando abarcar classes com características comuns. Por exemplo, a classe Evento-Científico dividiu-se em duas subclasses abstratas, Evento-Científico-ao-Vivo (com subclasses Conferência e Workshop) e Evento-de-Publicação-Científica (com subclasses Jornal e Revista). Esta mudança facilitou o reconhecimento e emprestou granularidade e coerência à ontologia.

Técnicas de PLN não foram empregadas nos protótipos. Com cada agente foram realizados três testes para classificação funcional e de conteúdo de páginas e dois testes para extração de informação. Para dois dos três primeiros testes, lançou-se mão de *corpus* de páginas recuperadas de consultas a mecanismos de busca; o primeiro *corpus*, para aquisição de conhecimento (definir casos, regras e *templates*) e o segundo para teste cego. O terceiro teste foi feito acessando diretamente a Web. Dois agentes do grupo científico foram elaborados: o agente CFP, que trata páginas de chamadas de trabalhos ("*Call for papers*") de eventos científicos, como conferências e jornais, classificando-as em oito classes de páginas (as quatro citadas acima, mais Evento-

Genérico-ao-Vivo, Evento-Genérico-de-Publicação e Edição-Especial-de-Jornal e Revista) e o agente de artigos científicos, que processa páginas de artigos e documentos científicos, refinando consultas aos mecanismos de busca com palavras-chave bastante comuns em artigos: *"abstract"*, *"keywords"*, *"introduction"*, *"conclusion"*, etc. Este último classifica as páginas em artigos de *workshop*, conferência, jornal e revista, capítulo de livro e artigos genéricos, além de teses, dissertações, relatórios técnicos e de projeto.

Cada uma das oito classes citadas anteriormente é representada por um conjunto de atributos. Confore colocado na seção 3.1, para cada atributo a ser extraído, foi criada uma instância de *template*. Como exemplo, considere o seguinte trecho de uma página CFP:

> "***Paper deadline***
> *Technical Paper must be* ***submitted by****:* ***July 17, 1995***
> *(Papers must be complete for review with all references, figures etc.);*
> ***Notification of acceptance****:* ***September 4, 1995*** *(Reviewers may suggest modifications.)*"

A data "17 de julho de 1995" (em inglês, "*July 17, 1995*") refere-se à palavra-chave "*submitted by*" que expressa uma data limite para entrega de trabalhos. Isto é definido pela proximidade entre os dados e a ausência de sinais que anulem esta relação, como é o caso da mesma data (17 de julho de 1995) e a palavra-chave "*Notification of Acceptance*", onde entre os dois dados extraídos ocorre o sinal de ponto-e-vírgula que anula a relação entre eles.

4.1. Resultados

Os resultados obtidos se referem a dois conjuntos de testes. O primeiro compreende um total de quatro testes para avaliar o desempenho do MASTER-Web na classificação funcional e de conteúdo. O segundo conjunto avalia o desempenho do sistema no processo de extração de informações.

4.1.1. Classificação Funcional e de Conteúdo

A figura 2a mostra as performances dos agentes. Nela, reconhecimento indica se uma agente identificou corretamente páginas-conteúdo. A classificação de páginas-conteúdo Um quarto teste foi rodado com o agente CFP na Web, desta feita beneficiando-se da categoria funcional listas. Listas de CFPs sobre um dado assunto costumam ser mantidas por pesquisadores e organizações.

Agente CFP: Mais de 70% dos eventos eram conferências. O agente reconheceu erradamente páginas longas, geralmente sobre um assunto ou de uma comunidade (linux, XML, etc). Chamadas de eventos que não empregam o jargão comum a eventos científicos não foram reconhecidas. Listas foram detectadas pela presença de um número factível de âncoras citando palavras-chave relativas a eventos, ou por uma certa quantidade de intervalos de tempo (como 1-4 de dezembro). As listas devem ser reconhecidas com precisão, pois podem levar ao tratamento de muitas páginas inúteis.

O uso de listas melhorou a recuperação de páginas úteis entre 13 a 22% (ver Figure 2b), retornando um conjunto mais focado. Páginas de outras classes, como

mensagens e sugestões para os agentes de Organizações e de Publicações-Divisíveis (que trata Anais e Livros) foram substituídas por páginas conteúdo sob a forma de *framesets*. Nos outros testes, só um *frame* foi encontrado.

	Agente "Call for Papers"				Agente Artigos		
	Corpus Aquisição	Corpus de Teste	Teste na Web	Web usando Listas	Corpus Aquisição	Corpus de Teste	Teste na Web
Reconhecimento	97,1	93,9	96,1	96,3	93,1	82,7	87,8
Categ Funcional	93,8	93,9	93,8	95,7	96,8	94,0	95,1
Classif.Conteúdo	94,9	93,3	92,9	91,7	97,0	93,0	81,4
Págs Processadas	244	147	129	188	190	150	184

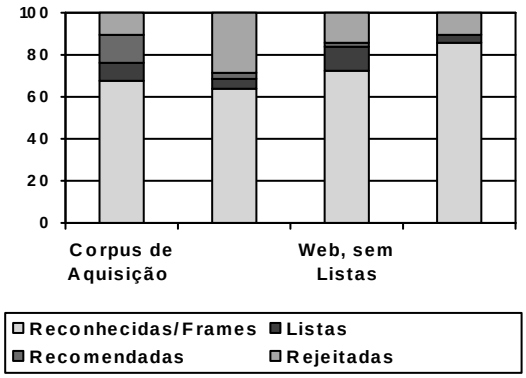

Figura 2. a) Performance dos agentes CFP e de artigos nos testes. b) Gráfico evidenciando o ganho de desempenho com o uso de listas no agente CFP.

Até o padrão das páginas rejeitadas mudou: ao invés das páginas erradas devidas à flata de precisão dos mecanismos de busca, vieram páginas gráficas iniciais de eventos. Digno de nota ainda que eventos encontrados a partir de listas não apareceram em outros testes. Estes fatos justificam a categorização funcional e o uso de listas.

Agente de artigos: Os erros no reconhecimento devem-se a artigos com poucos atributos, com atributos difíceis de identificar, - como a afiliação a uma empresa desconhecida - ou com atributos no fim do artigo. Um artigo deve possuir um conjunto mínimo destes atributos; artigos apenas com o nome do autor não foram reconhecidos, mas isto também ocorre em sistemas similares como o CiteSeer [Bollacker et al 98]. A classificação por conteúdo baseou-se em dados de publicação dos artigos, presentes no topo das páginas. Porém, mais da metade deles não traziam esses dados.

Cooperação: Efetuou-se um teste integrado, em que os agentes cooperaram. O agente CFP pediu ao agente de artigos âncoras no topo de artigos contendo conceitos como "conferência" e "jornal". Apesar de ter funcionado, apenas três páginas foram sugeridas pelo agente de artigos ao agente CFP e nenhuma página foi sugerida erradamente.

Para evidenciar que a cooperação pode ser útil, o agente CFP procurou, no atributo *Comitê de Programa*, sugestões de páginas ao futuro agente de pesquisadores. Nenhum dicionário de nomes ou técnica de extração foram empregados. O agente sugeriu 30 *links* corretos e 7 errados, um bom resultado, pois páginas de pesquisadores são menos estruturadas, portanto difíceis de ser recuperadas por mecanismos de busca.

4.1.2. Extração de Informação

Foram executados dois testes de extração de informação com o agente CFP. O primeiro foi realizado com um *corpus* de teste para aquisição de conhecimento, onde se fez ajustes (correção de instâncias e seus atributos) para maximizar a extração, enquanto que o segundo visou obter resultados para validação do desempenho do sistema sem que ajustes fossem permitidos. A figura 3 mostra uma tabela dos resultados obtidos nos testes. A cobertura é um índice quantitativo, ou seja, a porcentagem de informações

extraídas das páginas, enquanto que a precisão é um índice qualitativo, referindo-se ao número de informações que foram extraídas corretamente das páginas.

	Cobertura(%)	**Precisão(%)**
Local	68,75	68,75
Deadline	75,00	71,43
Período	78,49	57,89
Lista de Tópicos	70,00	60,00
Data de Aceitação	88,89	77,78
Total	75,60	67,06

Figura 3. Cobertura e precisão na extração de informação realizada pelo agente CFP.

As informações extraídas foram: local, *deadline* e lista de tópicos. Os resultados obtidos dependem diretamente das instâncias criadas na ontologia, ou seja, da experiência do especialista humano. A extração de informação atingiu um índice médio de cobertura de 75,6 e de precisão de 67,06, mostrando que a extração foi eficiente. É importante ressaltar que não foram usadas técnicas de aprendizado que aumentariam sensivelmente a precisão do sistema.

5. Comparação com trabalhos similares

5.1. WebKB: Classificação e Extração Baseadas em Aprendizado e Ontologias

O sistema WebKb [Craven et al 98] aprende automaticamente regras de categorização e extração integrada de páginas na Web, empregando uma ontologia do domínio com classes e relacionamentos, definida num formalismo que pode permitir inferência. As páginas da Web são representadas, com título, palavras-chave, freqüências e ponteiros.

A decisão de usar aprendizado automático depende de alguns fatores. O primeiro é uma comparação entre os custos de anotação de *corpi* e o trabalho de inspeção e aquisição do conhecimento. Existem vantagens em usar aprendizado, como velocidade e adaptabilidade, e desvantagens como legibilidade, engajamento ontológico das regras aprendidas – que tendem a ser muito específicas -, e dificuldades de aproveitar conhecimento *a priori*, de capturar regras sem introduzir porção de características para o aprendizado, e de generalizar sobre um grande número de características ou classes.

O sistema emprega uma ontologia do domínio com, apenas quatro entidades: atividades (e.g. projetos e cursos), pessoas (estudante, professor, membro do *staff*, etc), e departamentos. Relações também estão presentes, como instrutores de cursos, membros de projeto, orientadores, e outras.

Os autores do WebKB avaliam a classificação apenas através dos falsos positivos, reportando percentagens entre 73 e 83 %, exceto para as classes *membro do Staff* e *outros* (rejeitadas). Contudo, se computados os falsos negativos, a classe *outros* tem boa performance (93,6%), a classe estudante tem 43% e as outras seis classes comportam-se abaixo de 27%, baixando a média de acerto para apenas cerca de 50%. Isto leva à hipótese de que a ontologia empregada no WebKB não tenha sido abrangente o suficiente. Já a ontologia de Ciência usada pelo MASTERWeb possui classes, como projetos e produtos, que não foram usadas por dois motivos: os agentes precisam destes conceitos para suas funções, e futuros agentes que tratem delas podem ser elaborados.

Por outro lado, uma ontologia com muitas classes pode dificultar a generalização do aprendizado. Neste caso, seriam necessários mais agentes com aprendizado.

5.2. Os Sistemas CiteSeer e DEADLINER

Estes sistemas perfazem uma eficiente recuperação, filtragem e extração da Web, usando métodos estatísticos e de aprendizado combinados com conhecimento a priori.

O CiteSeer [Bollacker at al 99] é um dos mais usados na busca de artigos científicos. O sistema monitora newsgroups e editores e mecanismos de busca a partir dos termos “publications”, “papers” e “postscript”. São extraídos dados bibliográficos do artigo e da bibliografia, que atua como lista, ajudando a achar outros artigos.

O DEADLINER [Kruger et al 2000] busca anúncios de conferências, extraindo deles data inicial, final e limite, comitê, afiliação de membros do comitê, temas, nome do evento e país. A performance de reconhecimento do DEADLINER está acima de 95%, contudo, sua definição de evento é mais restritiva: todos os atributos têm de estar presentes, exceto país, além de dados de submissão. O MASTER-Web oferece mais flexibilidade e cobertura, aceitando anúncios de capítulos de livros, jornais, revistas e concursos. Os requisitos estão em casos, que são mais flexíveis.

O problema destes sistemas é que, mesmo sendo confeccionados pelo mesmo grupo de pesquisa, ambos deparam com *links* que interessam ao outro, e, não podem repassá-los. Sob o prisma de uma possível multiplicação de extratores pela Internet, isto deriva de um problema de representação de conhecimento: os dois sistemas (e outros que surgirão) não podem expressar intenções como pedir páginas ou sugerir *links*, pela falta de ontologias do domínio ou de páginas da Web. O conhecimento destes sistemas está escondido dentro de algoritmos, não sendo possível o compartilhamento deles para outros sistemas, nem a especificação de contextos em que seriam úteis. Outra vantagem está no reuso massivo da arquitetura, em que apenas parte do conhecimento tem de ser descoberto e especificado. Numa abordagem como a do CiteSeer e DEADLINER, para processar uma nova classe, um novo sistema precisa ser elaborado, sem maiores reusos.

6. Trabalhos futuros e conclusões

O projeto pode estender-se em várias direções. Novos agentes para o grupo serão desenvolvidos, como o agente de pesquisadores, e a cooperação tornar-se-á mais efetiva. Técnicas de aprendizado e PLN serão incluídas visando lidar com classes de páginas menos estruturadas. Alguma forma de checar duplicatas também será implementada.

Os agentes cognitivos, por basearem-se em modelos com conhecimento, podem comunicar-se e evitar redundância de tarefas num mesmo ambiente, Isto pode proporcionar a inauguração de uma nova era na informática distribuída, a comunicação em nível de conhecimento, dinamicamente estabelecida durante a execução, e o processamento de nichos de informação consistentes, como o domínio científico, o domínio turístico, etc. A idéia motivadora é a de que mecanismos de busca baseados em palavras-chave podem constituir a base para agentes cooperativos mais precisos e focados em domínios restritos, baseados em ontologias.

Bibliografia

L O Álvares, J S Sichman (1997). Introdução aos sistemas multiagentes. In C M B Medeiros, editor, *Jornada de Atualização em Informática (JAI'97)*, chapter 1, pages 1–38. UnB, Brasília.

R Benjamins, D Fensel and A G Pérez. (1998) Knowledge Management through Ontologies. Proc. of the 2nd International Conf. on Practical Aspects of Knowledge Management, Basel, Switzerland.

K Bollacker, S Lawrence and C L Giles. (1998) CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Proceedings of the 2nd International ACM Conference on Autonomous Agents, USA.

M Craven, A McCallum, D DiPasquo, T Mitchell, D Freitag, K Nigam. Stephen Slattery. (1999) Learning to Extract Symbolic Knowledge from the World Wide Web. Technical Report CMU-CS-98-122. School of Computer Science. CMU, USA.

F Freitas (2001) Ontology of Science. http://protege.stanford.edu/plugins/ontologyOfScience/ontology_of_science.htm

F Freitas, G Bittencourt (2003) An Ontology-based Architecture for Cooperative Information Agents. To appear in: Proceedings of International Joint Conference of Artificial Intelligence (IJCAI`2003), Acapulco, Mexico.

A Kruger, C L Giles, F Coetze, E Glover, G Flake, S Lawrence and C Omlin. (2000) DEADLINER: Building a new niche search engine.Conf. on Information and Knowledge Management, Washington DC.

G Miller (1995) WordNet: A Lexical Database for English. Communications of the ACM. 38(11):39-41. EUA.

E. Riloff. (1994) Information Extraction as a Basis for Portable Text Classification Systems. PhD. thesis. Dept of Computer Science. Univ of Mass. at Amherst. USA.