Ontologias detalhadas e classificação de texto: uma união promissora

Eunice Palmeira[1,2], Fred Freitas[3]

[1]Coordenadoria de Tecnologia e Informática – Centro Federal de Educação Tecnológica de Alagoas – UNED/PIn
Av. das Alagoas, s/n - Palmeira de Fora - 57601-220
Palmeira dos Índios – AL – Brasil

[2]Instituto de Computação - Universidade Federal de Alagoas
Campus A. C. Simões, BR 104 - Norte, Km 97, Cidade Universitária – 57072-970
Maceió – AL - Brasil

[3]Centro de Informática - Universidade Federal de Pernambuco
Av. Prof. Luis Freire, s/n, Cidade Universitária – 50740-540
Recife-PE - Brasil

eunice@cefet-al.br, fred@cin.ufpe.br

Abstract. *A lot of information may not be found in the Web ought to the lack of tools capable of treating semantically the information contained in documents. This work shows that ontologies can play a crucial role in cooperative information gathering (CIG) systems. The experiments have led us to draw two interesting conclusions: (1) the use of a thoroughly detailed ontology about a given domain assure promising results for classification, even with the application of simple techniques such as a small set of production rules; (2) an ontology-based CIG system can be easily carried for other domains, provided that a detailed ontology exists for this domain.*

Resumo. *Muita informação pode não ser encontrada na Web por falta de ferramentas capazes de tratar semanticamente a informação contida em documentos. Este trabalho mostra que ontologias podem ter um papel crucial num sistema de manipulação de informação. Os experimentos conduzem a duas conclusões interessantes: (1) uma ontologia bem detalhada a respeito de um dado domínio leva a resultados de classificação de texto bastante promissores, mesmo com a aplicação de técnicas simples, como um pequeno conjunto de regras de produção; (2) um sistema de manipulação de informação usando ontologias pode ser facilmente portado para outros domínios, desde que exista uma ontologia detalhada para o novo domínio.*

1. Introdução

A heterogeneidade e o enorme volume de informação disponível na Web constituem um fator complicador para quem busca informação específica e objetivamente útil. Deste problema, surge a necessidade de ferramentas eficazes na pesquisa de informação e capazes de tratar semanticamente a informação contida em documentos que foram escritos com uma estrutura preocupada apenas com a exibição de dados.

É também observada uma tendência para o desenvolvimento de programas capazes de executar automaticamente tarefas múltiplas relacionadas a recuperação de informação, tais como selecionar documentos de uma base, agrupá-los em categorias ou classes e deles extrair determinados conjuntos de informações [Jackson and Moulinier, 2002]. Contudo, algumas destas ferramentas atuam sobre domínios muito restritos, vinculadas a bases específicas, não possuindo flexibilidade para recuperação, classificação e extração de dados em diferentes campos de conhecimento.

Acredita-se que o uso de ontologias aliado a sistemas capazes de derivar conclusões e inferir sobre a relevância da informação pode fornecer uma compreensão mais apurada do domínio, resultando numa classificação textual eficiente. Essa união de ontologias e sistemas baseados em conhecimento favorece a manipulação de informação – entendida como a realização de mais de uma tarefa, como por exemplo, classificação, recuperação e extração, proporcionando flexibilidade às ferramentas para atuar em diferentes domínios. Observa-se ainda a vantagem de que para trocar de domínio basta trocar a ontologia e parte da base de regras.

Para comprovar as hipóteses de pesquisa enumeradas acima, foram realizadas as seguintes atividades:

- construção de uma ontologia de Inteligência Artificial (IA) com alguns dos ramos (e.g. Busca) bastante detalhados, que está descrita na seção seguinte;
- alteração do sistema de manipulação de informação MASTER-Web para que ele, além de recuperar, classificar e extrair atributos, trabalhe com artigos científicos, i.e., reconheça e extraia seções desses artigos;
- desenvolvimento de heurísticas de classificação de artigos, inserindo-as em regras de produção manipuláveis pelo MASTER-Web. O sistema e as alterações efetuadas, bem como as heurísticas e suas respectivas regras, estão comentadas na seção 3;
- realização de um teste de aquisição de conhecimento e um teste com um corpus desconhecido para avaliar a metodologia de classificação de artigos científicos aplicada neste trabalho. As condições da base experimental, os resultados obtidos e sua discussão encontram-se na seção 4;

Por fim, são feitas as considerações finais e relatados os trabalhos futuros na seção 5.

2. Ontologias e seus usos para manipulação de texto

A descrição do trabalho inicia-se com um breve conceito sobre ontologias, seus usos e benefícios, seguido da apresentação da ontologia de IA aqui utilizada.

Segundo [Guarino, 1998], ontologias como um ramo da Filosofia se refere a um sistema de categorização dos objetos do mundo para a organização da realidade. Assim, ontologias especificam conceitos sobre um dado campo de conhecimento e as relações, restrições e axiomas válidos entre esses conceitos e suas instâncias, provendo uma boa taxonomia tanto de entidades como de instâncias deste universo de discurso.

Atualmente, o uso de ontologias vem se difundindo em diversas áreas, como na Web Semântica [Berners-Lee et al., 2001], Direito [Lame, 2000], Comércio Eletrônico [Morgenstern and Riecken, 2005] entre outras áreas que buscam desenvolver um vocabulário contendo os conceitos relativos ao domínio de aplicação.

2.1 Ontologia para o domínio de Inteligência Artificial

Este trabalho apresenta uma ontologia que contém informações sobre os conceitos investigados no domínio de Inteligência Artificial (IA), seus ramos, subáreas, teorias, técnicas, métodos e aplicações, Esta ontologia constituirá um elemento crucial para a classificação de textos sobre IA, como será visto nas seções restantes.

O propósito da ontologia de IA apresentada nesse trabalho é servir como repositório semântico para a representação do significado de seus termos. Esta ontologia pode ser empregada em sistemas baseados em conhecimento, e, em especial, para sistemas de manipulação cooperativa de informação, como sistemas extratores ou classificadores de informação, compondo uma base de conhecimento e servindo como recurso de apoio à informação. Além do mais, uma ontologia como esta pode servir como vocabulário na comunicação entre agentes em cooperação.

A IA abrange uma enorme variedade de subcampos, desde áreas de uso geral, como aprendizado e percepção, até tarefas específicas como jogos de xadrez, demonstração de teoremas matemáticos e diagnósticos de doenças [Russell and Norvig, 1998]. A própria dimensão e as diversas relações existentes entre os problemas, métodos, técnicas, etc. contribuem para estabelecer a falta de concordância da comunidade de pesquisa científica quanto às várias ramificações da área, considerando as sutis interseções e interações envolvidas. Alguns autores apresentam discordâncias quanto à classificação dos ramos, enquanto outras subáreas estão estruturadas e solidificadas, como é o caso da Busca.

Foram desenvolvidas separadamente ontologias de subáreas e aplicações que posteriormente são agrupadas para compor a ontologia de IA. O escopo dessa ontologia compreende, a princípio, as áreas de Busca, Representação do Conhecimento, Redes Neurais e Aprendizado de Máquina. Ontologias de outras subáreas de IA poderão ser incorporadas estendendo a ontologia existente.

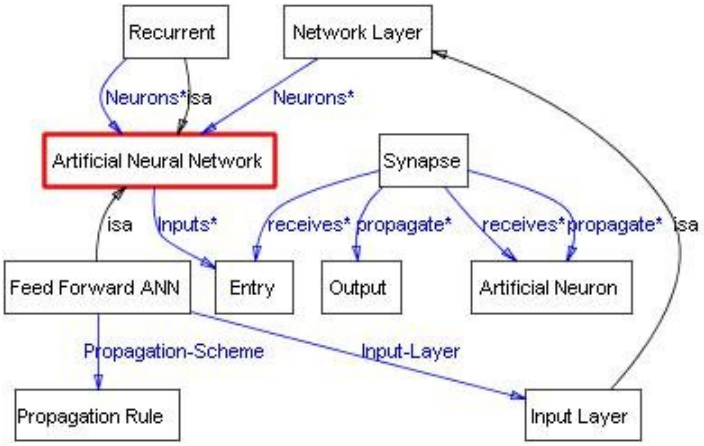

Figura 1. Algumas classes e seus relacionamentos no domínio de Redes Neurais.

Na figura 1, destacam-se alguns relacionamentos entre as principais classes definidas na ontologia de Redes Neurais. Dois tipos de Redes Neurais, *Recurrent* e *Feed Forward ANN*, bem como alguns conceitos relevantes, como *Synapse*, *Neuron e Layer*, são apresentadas nessa figura. Várias relações estão disponíveis entre estes conceitos.

As ontologias dos ramos de Busca, Representação do Conhecimento e Aprendizado de Máquina foram desenvolvidas com um nível de detalhes semelhante à de Redes Neurais.

3. O uso do MASTER-Web para classificação textual

3.1. O sistema MASTER-Web

O MASTER-Web (*Multi-Agent System for Text Extraction, Classification and Retrieval over the Web*) é uma arquitetura de Sistemas Multiagentes Cognitivos para resolver o problema da extração integrada de entidades pertencentes às classes que integram um grupo de páginas ou (*cluster*) [Freitas and Bittencourt, 2003]. Este sistema apresentou bons resultados nas tarefas de recuperação, classificação e extração de informação, e permite a cooperação entre agentes para a realização destas tarefas. Utiliza uma abordagem baseada em conhecimento e apresenta vários tipos de reuso, como agentes compartilhando da mesma estrutura em termos de código, serviços dos mecanismos de busca, e boa parte do conhecimento de que dispõem (ontologias e regras de produção), o que facilita e acelera a construção de novos agentes.

Cada agente, representado na figura 2, reconhece, filtra e classifica páginas que correspondem a instâncias da classe que ele processa (por exemplo, páginas de pesquisadores, chamadas de eventos científicos - 'Call for Papers'), extraindo também seus atributos (por exemplo, áreas de pesquisa e instituição dos pesquisadores). Cada agente possui ainda um meta-robô, que se conecta a múltiplos mecanismos de busca - como Google, Excite e outros. Ele consulta os mecanismos de busca com palavras-chave que garantem cobertura[1] em relação à classe de páginas processada pelo agente. (e.g., os termos 'call for papers' e 'call for participation' para o agente CFP).

Como pode ser observado na figura 2, cada agente desempenha quatro tarefas consecutivas no processamento de cada página [Freitas and Bittencourt, 2003]: validação, pré-processamento, reconhecimento e extração.

A **Validação** elimina as páginas inacessíveis, repetidas ou em formatos que os agentes não possam processar. A fase de **Pré-processamento** tem por meta representar as páginas de várias maneiras, tais como conteúdo com e sem HTML, palavras-chave e suas freqüências, dentre outras, com dados extraídos delas, aplicando, se necessário, recuperação de informação e processamento de linguagem natural. Esses dados são então passados ao motor de inferência. Nas fases de **Classificação** e **Extração de Atributos,** o sistema descobre se a página é do domínio tratado, reconhece de que classe a página tratada é instância e extrai atributos que irão compor a instância da classe.

3.2 MASTER-Web para classificação de artigos científicos

Este trabalho visa provar a portabilidade de uso de sistemas baseados em conhecimento para tarefas relacionadas a Recuperação de Informação, e, mais especificamente, do sistema MASTER-Web. Portanto, é efetuada a classificação de artigos científicos, identificando os das áreas de Inteligência Artificial contidas na ontologia utilizada. Para

[1] Cobertura (recall) significa o quociente entre o total de documentos relevantes recuperados sobre o total de documentos relevantes.

tanto, fez-se necessário alterar o sistema estendendo a representação das páginas HTML existente no MASTER-Web com uma representação das seções dos artigos científicos.

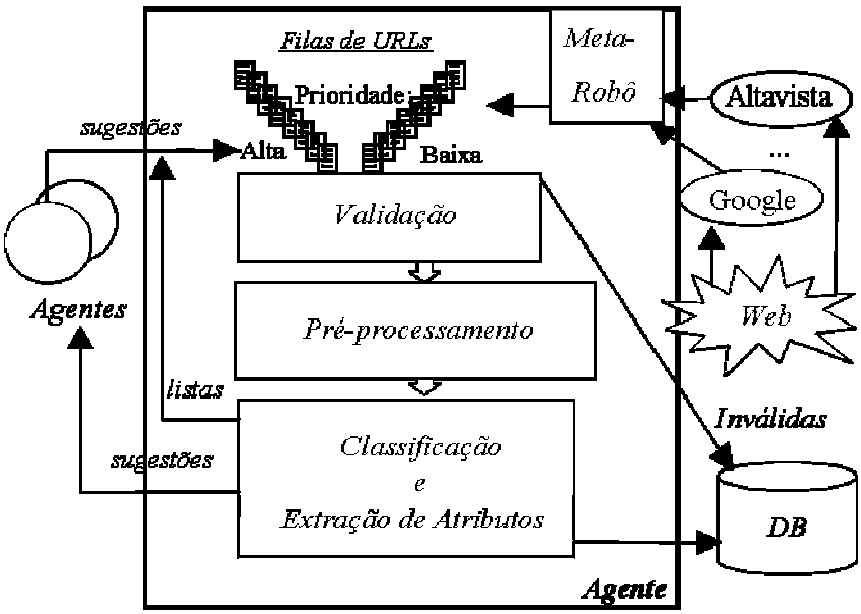

Figura 2. Detalhe de um agente, mostrando as tarefas a serem realizadas [Freitas and Bittencourt, 2003].

Assim, na fase de pré-processamento, também foi incluído o reconhecimento de alguns tipos específicos de seções de artigo (e.g. introdução, trabalhos relacionados, conclusões, etc). Essas seções são classificadas quanto à sua relevância como segue:

• **Seção irrelevante:** Seu conteúdo aborda assuntos gerais, citações das obras de referências consultadas ou sugestões de trabalhos a serem realizados futuramente. Alguns exemplos de seções deste tipo são: introdução, bibliografia, trabalhos relacionados, trabalhos futuros, apêndices, entre outros. A conclusão é que este tipo de seção, que não trata diretamente sobre o tema do artigo, pode deteriorar a precisão dos resultados de classificação, devendo portanto ser ignorada.

• **Seção relevante:** Seção que trata especificamente sobre o tema do artigo, de modo a apresentar os aspectos teóricos, e por isso, é considerada significante para classificação. Resumo, palavras-chave e conclusão são exemplos desse tipo de seção.

3.3 Estratégias de classificação

O processo de classificação de artigos quanto aos assuntos abordados tem início com o reconhecimento de suas seções e dos subdomínios de IA. As fases desse processo de classificação são as seguintes:

• **Pré-processamento com Reconhecimento de seções.** A tarefa é analisar o artigo reconhecendo seções relevantes, identificando e extraindo delas os termos presentes na ontologia do domínio.

• **Reconhecimento de Classes.** Dentro do domínio de IA existem subdomínios, que são chamados de classes principais, ou seja, classes que representam assuntos do domínio (e.g. Engenharia do Conhecimento, Ontologias, etc). As classes principais são explicitamente definidas como classes a serem reconhecidas, devendo suas subclasses

serem reconhecidas quando presentes no texto, o que torna a classificação mais específica. Para isso foram criadas regras conforme as seguintes descrições:

- *Reconhecimento Direto de Classes Principais***.** O reconhecimento das classes principais é efetuado quando os nomes destas classes, de suas subclasses ou ainda seus sinônimos encontram-se presentes nas seções do artigo. No caso específico de IA, temos como exemplo de classes principais *Artificial Neural Networks*, *Knowledge Acquisition*, *Knowledge Representation Formalisms*, *Machine Learning* entre outras.

- *Reconhecimento de Classes Através de Atributos***.** Quando duas classes estão ligadas por meio de um atributo, essas ligações podem ser consideradas fortes ou fracas, dependendo do vínculo de dependência ou interação entre as partes. Por exemplo, a relação entre *Output Layer* e *Artificial Neural Network* pelo atributo *Neurons*, é uma relação forte, enquanto que a relação *Inputs* entre as classes *Artificial Neural Network* e *Entry* é considerada uma relação fraca porque o termo *Inputs* não é tão discriminante quanto *Neurons*. Para que o sistema reconheça bem as classes, as ligações fortes devem ser explicitamente definidas dentro da ontologia. Tendo posto essas relações, é possível reconhecer classes principais ao encontrar no texto classes que possuem esse tipo de ligação com as classes principais.

- *Reconhecimento de Classes Através de Relação Indireta***.** Algumas classes não estão ligadas por meio de um relacionamento, mas por meio de outras classes. Por exemplo, a classe *Knowledge Level* está associada à classe *Ontology* que por sua vez está relacionada, pelo atributo *formalism*, a *Domain-oriented KR Formalism* que é subclasse de *Knowledge Representation Formalisms*. Nesse caso, a classe *Knowledge Level* possui uma relação indireta com as classes principais *Knowledge Representation Formalisms* e *Domain-oriented KR Formalism*. Portanto, classes principais podem ser reconhecidas quando classes que possuem esse tipo de relação com as classes principais são identificadas no texto.

• **Ordenação e Classificação.** O objetivo é aumentar a precisão da classificação, considerando a freqüência dos termos no texto e o peso das seções onde os termos são encontrados. Neste último caso, apenas a seção resumo possui um peso diferente das demais seções, por se tratar de uma síntese do artigo. Adotou-se um esquema de pontuação onde as classes principais reconhecidas são pontuadas de modo a ordená-las e então é efetuada a classificação do artigo quanto aos assuntos abordados. A pontuação se dá pelos quatro itens abaixo:

- *Sinônimos***:** Esse item é pontuado com a freqüência total dos termos encontrados no artigo referentes à classe principal em questão. Isso inclui o número de sinônimos encontrados da classe principal, das suas subclasses e das classes relacionadas direta e indiretamente a classe principal, como mostra a fórmula abaixo.

$$Sinônimos = \sum freqüência\ de\ cada\ termo\ reconhecido\ referente\ a\ classe\ principal$$

A tabela 1 mostra um exemplo para a pontuação dos itens referente à classe principal *Knowledge Representation Formalisms* para um artigo. Neste caso, o item *Sinônimos* é igual a 9.

Tabela 1: Exemplo da pontuação referente a classe principal *Knowledge Representation Formalisms*.

Reconhecimento	Sinônimos	Freqüência	Seções	Seção Especial
Direto (Classes)	KR Formalism	1	Other-sections	0
Direto (Subclasses)	Semantic Net	6	(other-sections, conclusion)	0
Por atributo	----------			0
Relação indireta	Ontology	2	Other-sections	0
∑ = 3		**∑ = 9**	**∑ = 4**	**∑ = 0**

- *Tipos de reconhecimento*: Classes principais podem ser reconhecidas diretamente (através de classes e subclasses) através de atributos e/ou por relação indireta. Nesse caso, a pontuação é definida pelo somatório dos tipos de reconhecimento que uma classe principal foi reconhecida. Quanto maior esse valor, maior a certeza da classificação. Para o exemplo apresentado pela tabela 1, a pontuação desse item é igual a 3, dado que dos quatro tipos de reconhecimento apenas o reconhecimento por atributo não reconheceu nenhum termo.

$$Tipos\ de\ reconhecimento = \sum tipos\ de\ reconhecimento\ da\ classe\ principal$$

- *Tipos de seção*: Os termos do domínio podem ser encontrados em várias seções do artigo. Por conseguinte, o valor desse item é igual ao somatório do número total de seções onde cada termo é referenciado, como definido na fórmula abaixo.

$$Tipos\ de\ seção = \sum número\ de\ seções\ onde\ cada\ termo\ é\ reconhecido$$

Considerando o exemplo representado pela tabela 1, o valor deste item é igual a 4, pois o termo *KR Formalism* é encontrado na seção *other-sections*. Já *Semantic Net* aparece tanto na seção *other-sections* quanto em *conclusion* e o termo Ontology aparece apenas na seção *other-sections*.

- *Seção especial*: Se o artigo possui a seção resumo, este item recebe como pontuação o número de vezes que os termos forem referenciados nesta seção.

$$Seção\ especial = \sum freqüência\ de\ cada\ termo\ reconhecido\ na\ seção\ resumo$$

Os valores fornecidos em cada um dos itens serão utilizados para determinar a ordenação das classes principais. A ordenação é iniciada pelo item *Sinônimos* seguido por *Tipos de reconhecimento*, *Tipos de seção* e por fim o item *Seção especial*.

A tabela abaixo é um exemplo do resultado da ordenação das classes principais reconhecidas em um determinado artigo:

Tabela 2: Exemplo da ordenação das classes principais reconhecidas em um artigo.

Classe Principal	Sinônimos	Tipos de reconhecimento	Tipos de seção	Seção especial
Knowledge Representation Formalisms	4	3	2	1
Ontology	37	1	2	0
Search	1	1	0	0

A área atribuída ao artigo é inferida a partir dos valores dos quatro itens, obedecendo a pelo menos uma das seguintes premissas:

• Sinônimos > 2; ou Tipos de reconhecimento. > 1; ou Tipos de seção >; 1 ou Seção especial >= 1.

Pela análise de artigos na fase de aquisição de conhecimento, os valores acima foram considerados como boas heurísticas para classificação. Além disso, as classes principais foram privilegiadas para garantir um bom ordenamento.

Analisando o exemplo da tabela 2, apenas *Search* não é atribuído como assunto tratado pelo artigo. De acordo com as três primeiras condições acima, a ausência de termos relacionados à classe *Search* na seção resumo e a insatisfação dos valores das demais categorias declaram que o artigo apenas cita a área *Search* e fala sobre *Knowledge Representation Formalisms* e mais especificamente sobre *Ontology*.

4. Experimentos

Foram realizados um teste para aquisição de conhecimento e um teste com um corpus desconhecido, utilizando como dados experimentais uma base de texto composta por artigos do domínio de IA e também de outros domínios, como outros ramos da Computação, Medicina, Biologia, Economia, Filosofia, etc.

Os artigos que formam a base de textos são, ainda, heterogêneos quanto a divisão em seções. Alguns artigos apresentam o texto dividido em diferentes seções e subseções, por exemplo, resumo, sumário, introdução, trabalhos relacionados e bibliografia, enquanto outros possuem uma divisão menos preocupada com as diversas partes de construção do texto e consideram basicamente duas seções, a seção principal, de que trata o assunto do artigo, e a bibliografia. Observa-se que o sistema deve apresentar robustez frente a esta falta de padronização de seções dos artigos.

A base experimental foi composta por documentos em HTML, obtidos em diversos sites de universidades, professores e revistas. Os experimentos realizados classificaram os artigos com base na análise das seções do texto consideradas relevantes e no domínio ontológico. Cada artigo foi classificado como sendo ou não do domínio em questão, no caso IA, atribuindo aos artigos as áreas de abrangência.

4.1 Resultados

Os resultados obtidos apresentaram promissores percentuais de desempenho do sistema, como mostra a tabela abaixo. Para analisar os resultados apresentados, é preciso levar em conta o escopo da ontologia do domínio investigado, que compreende apenas as áreas de Busca, Representação do Conhecimento, Redes Neurais e Aprendizado de Máquina. Os artigos da área de IA, que não tratam de áreas compreendidas na ontologia, naturalmente não são classificados dentro deste domínio. Nesse contexto, a classificação garante um reconhecimento baseado na especificidade dos conceitos ontológicos, restringindo o domínio às classes e relações explicitamente definidas na ontologia. Outra observação digna de nota é que um artigo pode abranger mais de uma área, como por exemplo, falar tanto de *Machine Learning* quanto de *Search*. A tabela 3 mostra os resultados da classificação dos artigos por área de reconhecimento.

Foi realizado um outro experimento utilizando como base de testes os artigos classificados pelo MASTER-Web como sendo do domínio de IA. Neste experimento foi analisada apenas a seção resumo, desprezando assim as demais seções. Os resultados obtidos revelaram baixo percentual de acerto, em torno de 50% apenas. Analisando esses artigos, foi observado que o conteúdo dessa seção é em sua maioria abstrato, não retratando os temas abordados pelo artigo e dificultando a classificação da informação por parte do sistema, sobretudo a classificação em um menor nível de granularidade.

Tabela 3: Percentuais de acerto de classificação de artigos por área de reconhecimento.

Reconhecimento	Corretos	Falsos Positivos	Falsos Negativos	Acerto (%)
Artificial Neural Network	48	1	2	94,1
Knowledge Acquisition	17	0	1	94,4
Knowledge Engineering	3	0	0	100,0
Knowledge Representation Formalisms	56	9	1	84,8
Machine Learning	51	2	6	86,4
Ontology	19	0	0	100,0
Search	38	1	1	95,0
Outros domínios	228	7	11	92,7

4.2. Discussão

Os resultados apresentados mostram que a ferramenta apresentou um desempenho bastante significativo. O sucesso dessa classificação deve-se ao uso de uma boa ontologia, uma vez que melhora a precisão dos resultados, mesmo quando aliada à aplicação de técnicas simples, como uma pequena base de regras de produção. Isso se deve ao fato de que as relações detalhadas entre os diversos conceitos da ontologia favorecem a inferência correta do sistema durante a classificação, pois pela ligação de termos distintos de várias formas pode-se chegar a um novo conceito, como descrito na seção 3, item Reconhecimento de classes. Então, quanto mais detalhada a ontologia, maior será a especificidade da classificação. Outro aspecto importante diz respeito a portabilidade do sistema para manipulação de informação usando ontologias, permitindo a classificação em diferentes áreas pela simples troca do domínio ontológico.

A escolha das seções do artigo também possui um papel importante, onde seções que não abordam especificamente o tema são ignoradas, enquanto as seções que descrevem técnicas e comunicam resultados são consideradas e delas são extraídos os termos do domínio.

Alguns erros no reconhecimento das seções implicaram em falhas na classificação dos artigos. Esses erros foram ocasionados pela ausência de tags que destaquem seções como bibliografia e outras. Logo seções irrelevantes não foram reconhecidas e por conseqüência foram misturadas às seções relevantes, atrapalhando a classificação nestes (raros) casos. Outro fator de erro é a falta de termos usados por alguns artigos na lista de sinônimos da ontologia, o que evidencia a pluralidade de conceitos na área de IA.

É possível concluir ainda, que, à medida em que o vocabulário do domínio é ampliado e a ontologia é refinada, a classificação tornar-se-á ainda mais precisa.

5. Conclusões e trabalhos futuros

Este trabalho apresenta uma abordagem para classificação de informação de textos de domínios específicos, demonstrando que através da aplicação de técnicas simples e uma ontologia bem formada pode-se chegar a resultados de classificação bastante promissores. Mais que isso, esta abordagem abre ainda possibilidades para classificação textual de novos domínios pela simples troca ou adição de ontologias sobre estes novos domínios, sem exigir praticamente nenhuma alteração no sistema e sem se preocupar em estabelecer padrões de estruturação dos documentos dispostos na Internet.

Como prova de conceito, foi construída uma ontologia do domínio de Inteligência Artificial e adotado o sistema MASTER-Web para a tarefa de classificação de artigos científicos utilizando regras de produção. O sistema sofreu apenas pequenas alterações para lidar com seções de artigos científicos. A combinação entre uma boa ontologia e regras simples conduziu a resultados bastante significativos: por exemplo, o sistema é capaz de identificar no texto as subáreas de IA que ele aborda e deriva conclusões corretas com alto grau de acerto, distinguindo bem os assuntos tratados pelo artigo daqueles que são brevemente citados no texto.

Trabalhos futuros estão previstos para a melhoria da classificação textual. O principal deles é comparar a abordagem baseada em uma ontologia detalhada com outros algoritmos de classificação bem difundidos (e.g. naive bayes, redes neurais, ID3, etc) [Jackson and Moulinier, 2002]..Uma outra tarefa relevante é a extensão da ontologia de IA, pela adição de ontologias que cubram subáreas como Agentes Inteligentes, Robótica, Visão de Máquina, entre outras. Assim, novas áreas de IA poderão ser reconhecidas nos textos.

O reconhecimento de outros tipos de formatos de artigos também pode ser adicionado ao sistema, como arquivos em formato pdf, doc, entre outros. Uma adição significativa é habilitar o sistema a classificar diversos tipos de textos, como e-mails, entrevistas, anúncios, trechos de texto e não apenas artigos científicos. Essa nova funcionalidade é perfeitamente possível e viabilizada pelo reuso de muitas regras existentes, que por serem gerais facilitam esse tipo de extensão no sistema.

Também estão previstas melhorias nas regras que ordenam a classificação dos artigos, fornecendo ao usuário um ranking baseado em percentual de certeza, agilizando a localização da informação necessária ao usuário. Melhorias na integração do agente de classificação de artigos com os demais agentes do MASTER-Web podem ser realizadas, promovendo a cooperação entre os agentes como propõe a arquitetura desse sistema.

Referências

[Berners-Lee et al., 2001] T.im Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 2001.

[Freitas and Bittencourt, 2003] F. Freitas. G. Bittencourt. An Ontology-based Architecture for Cooperative Information Agents. In: International Joint Conference of Artificial Intelligence (IJCAI), 2003.

[Guarino, 1998] N.Guarino. Formal ontology and information systems. 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, 1998.

[Jackson and Moulinier, 2002] Peter Jackson and Isabelle Moulinier. Natural Language Processing for Online Applications; Text Retrieval, Extraction and Categorization. John Benjamins, Amsterdam, Philadelphia, 2002.

[Lame, 2000] Guiraude Lame. Knowledge acquisition from texts towards an ontology of french law. EKAW'2000 Workshop on ontologies and texts, 2000.

[Morgenstern and Riecken, 2005] L. Morgenstern and D. Riecken. Snap: An action-based ontology for e-commerce reasoning. IBM T.J. Watson Research Center, 2005.

[Russell and Norvig, 1998] Stuart J. Russell and Peter Norvig. Artificial Intelligence. Prentice-Hall, 2 edition, 1998.