Recuperação de Documentos Jurídicos Baseada em um Tesauro

Berthier Ribeiro Neto
berthier@dcc.ufmg.br

Rodrigo Tôrres Assumpção
contato@integradorjuridico.com.br

Universidade Federal de Minas Gerais
30.123-970 Belo Horizonte-MG, Brazil

Resumo

Os métodos de recuperação de informação em bases textuais, largamente utilizados pelas máquinas de busca, são baseados em técnicas voltadas para coleções de documentos genéricos. Em domínios específicos, como o jurídico, a aplicação direta destes métodos leva a resultados de qualidade menor que a esperada. A razão básica é que tais métodos não consideram informação semântica associada ao domínio em questão. Particularmente, no caso do Direito Brasileiro, discutido neste artigo, informação semântica pode ser obtida do tesauro elaborado pelo Conselho de Justiça Federal (CJF) e da estrutura do documento jurídico. Neste trabalho, exploramos a utilização deste tesauro em um modelo de recuperação de documentos jurídicos. Através de experimentação, mostramos que este modelo leva a melhor precisão (qualidade das respostas) que modelos voltados para coleções genéricas.

Abstract

The methods of information retrieval in textual bases, broadly used by the search machines, are based on techniques returned for collections of generic documents. In specific domains, as the juridical, the direct application of these methods takes the results of smaller quality than the expected. The basic reason is that such methods don't consider semantic information associated to the domain in subject. Particularly, in the case of the Brazilian Right, discussed in this article, semantic information can be obtained of the thesaurus elaborated by the Conselho de Justiça Federal (CJF) and of the structure of the juridical document. In this work, we explored the use of this thesaurus in a juridical document retrieval model. Through experimentation, we showed that this model takes the best precision (quality of the answers) that models returned for generic collections.

1. Introdução

O modelo vetorial[12] de recuperação de informação é um dos modelos mais populares entre a comunidade científica para busca de informação em coleções genéricas. O modelo considera um espaço multidimensional onde a consulta e cada um dos documentos são representados por vetores distintos neste espaço. Cada uma das dimensões espaciais representa um termo distinto encontrado em qualquer dos documentos da coleção.

A similaridade entre um documento d_j e a consulta q é quantificada pelo coseno do ângulo entre o vetor consulta e o vetor que representa o documento. Apesar de sua simplicidade, o modelo vetorial funciona bem com soluções genéricas e se constitui em um dos pilares fundamentais da maioria dos sistemas de recuperação de informação modernos, incluindo as máquinas de busca na *Web*.

Em coleções específicas, como uma coleção de documentos jurídicos, o modelo vetorial gera resultados de qualidade inferior à esperada. A razão fundamental é que o modelo não engloba

informação semântica adicional que se encontra normalmente disponível. Por exemplo, no domínio jurídico, informação semântica de relevância pode ser obtida de um tesauro jurídico e da estrutura dos documentos jurídicos.

Neste trabalho, utilizamos o arcabouço de uma rede Bayesiana para combinar evidência gerada pelo modelo vetorial clássico com evidência obtida de um tesauro jurídico e da estrutura dos documentos jurídicos. O novo modelo é então avaliado experimentalmente. Nossos resultados indicam que melhoria considerável na qualidade das respostas (isto é, em sua precisão) pode ser obtida. O artigo está organizado como se segue. Primeiro, descrevemos as características e a sintaxe de construção do tesauro jurídico utilizado em nossos exemplos. A seguir, apresentamos diferentes formas de melhor contextualizar a consulta do usuário (expandindo-a com termos correlatos) a partir do tesauro jurídico. Na seção 4, apresentamos a estrutura dos documentos sobre jurisprudências no cenário da justiça brasileira. A seção 5 descreve o modelo vetorial tradicional tomado como referência neste trabalho. A seção 6 descreve o modelo Bayesiano de recuperação de informação e o arcabouço para a representação do problema jurídico. A seção 7 apresenta todas as etapas de experimentação e

discute os resultados obtidos. A seção 8 apresenta nossas conclusões.

2. Tesauro Jurídico do CJF

O Tesauro Jurídico do CJF[3] contém 7840 conceitos definidos em uma lista alfabética que abrange todas as áreas do Direito Brasileiro.

Construído por uma equipe de profissionais especializados na área de Direito, o tesauro possui uma estrutura e sintaxe bem definidos, que possibilitam representar de modo descritivo o inter-relacionamento dos diversos conceitos jurídicos. Estes conceitos, representados no tesauro por descritores, são listados em ordem alfabética, onde cada descritor/conceito sempre é definido de acordo com um dos dois modelos abaixo:

<NOME-DESCRITOR-A>
UP *<NOME-DESCRITOR-B>*
TG1 *<NOME-DESCRITOR-GENÉRICO>*
TGm *<NOME-DESCRITOR-GENÉRICO>*
TE1 *<NOME-DESCRITOR-ESPECÍFICO >*
TEn *<NOME-DESCRITOR-ESPECÍFICO>*
TR *<NOME-DESCRITOR-RELACIONADO>*

Modelo 1 de definição de descritores do tesauro do CJF

<NOME-DESCRITOR-B>
USE *<NOME-DESCRITOR-A>*

Modelo 2 de definição de descritores do tesauro do CJF

Em nossa notação, nomes de descritores, tais como *<NOME-DESCRITOR-A>* e *<NOMEDESCRITOR-B>*, estão escritos em itálico, enquanto os operadores de relação são escritos em negrito (**USE**, **UP**, **TR**, **TG1**, **TGm**, **TE1**, **TEn**). Note que utilizamos os termos *descritor* e *conceito* com o mesmo significado.

Os operadores **USE** e **UP** são utilizados para indicar os sinônimos, ou equivalentes, de um conceito. Como o tesauro do CJF tem por objetivo a padronização da nomenclatura jurídica, então um dos descritores sinônimos é sugerido como o mais indicado para a categorização de documentos jurídicos.

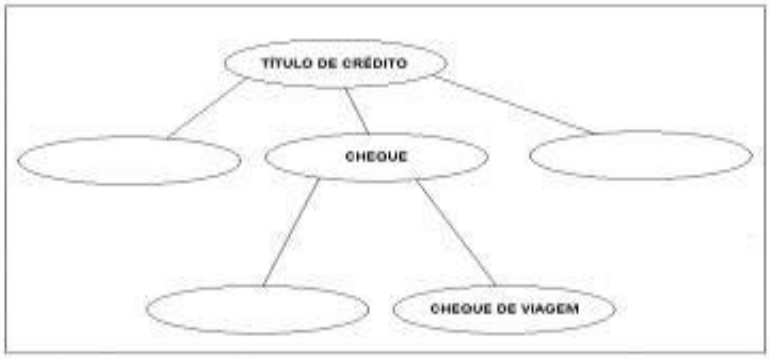
Em uma estrutura hierárquica de conceitos jurídicos, os operadores **TE** e **TG** indicam a especificidade e a generalidade. O operador **TG** (termo genérico) indica um conceito mais abrangente, do qual o termo subordinado (**TE** – termo específico) é um tipo. A cadeia hierárquica entre os conceitos é determinada pelos operadores de relação **TG** e **TE.** Os

operadores de relação **TG1** e **TE1**, por exemplo, determinam o descritor do primeiro nível hierárquico superior e também o primeiro nível hierárquico inferior respectivamente. E, consequentemente, **TG2** e **TE2**, determinam o segundo nível hierárquico.

É importante ressaltar que as definições dos conceitos jurídicos sempre devem refletir a realidade do mundo real. Portanto, um conceito pode fazer parte não apenas de uma única estrutura hierárquica, mas de diversas, o que permitiria a formação de uma polihierarquia.

O operador **TR** define uma relação associativa entre dois conceitos do tesauro que não são equivalentes nem formam uma hierarquia, mas que são semanticamente relacionados.

A figura 1 abaixo ilustra uma pequena porção do tesauro do CJF associada ao conceito *'CHEQUE DE VIAGEM'*. Um '*CHEQUE DE VIAGEM'* é um termo específico associado a um termo mais genérico que é o *'CHEQUE'* . Este, por sua vez, é um termo específico associado a um termo mais genérico que é o título de crédito. Os termos *'CHEQUE ADMINISTRATIVO'* e '*VIAGEM'* são termos relacionados ao *'CHEQUE DE VIAGEM'*.



CHEQUE DE VIAGEM
UP *TRAVELLERS CHECK*
TG1 *CHEQUE*
TG2 *TÍTULO DE CRÉDITO*
TR *CHEQUE ADMINISTRATIVO*
TR *VIAGEM*

TRAVELLERS CHECK
USE *CHEQUE DE VIAGEM*

Figura 1- Exemplo de estrutura conceitual hierárquica do tesauro do CJF.

Ademais, *'CHEQUE DE VIAGEM'* e *'TRAVELLERS CHECK'* são conceitos sinônimos. Porém, o conceito mais indicado para a utilização na categorização de documentos jurídicos é *'CHEQUE DE VIAGEM'*(isto é, se *'TRAVELLERS CHECK'* **USE** *'CHEQUE DE VIAGEM')*. A fim de se evitar repetições no tesauro, convencionou-se que apenas um dos conceitos conterá a definição completa do conceito jurídico, no caso, aquele que precede o operador **UP**, ou seja, *'CHEQUE DE VIAGEM'* .

3. Expansão da consulta

Expansão da consulta é uma técnica comumente utilizada na área de recuperação de informação para melhor contextualizar a consulta original do usuário. Isto é feito adicionando-se termos correlatos aos termos originalmente incluídos na consulta.

Neste trabalho, utilizamos o tesauro jurídico do Conselho de Justiça Federal (CJF) para adicionar à consulta original termos e expressões do tesauro que estão semanticamente relacionados ao conceito jurídico expresso pela consulta. Em nossos experimentos, cada consulta é representada por um termo ou conjunto de termos que definem um único conceito jurídico definido no tesauro. Cada conceito do tesauro do CJF é definido por outros conceitos que também são definidos no tesauro. A partir desta característica, foram desenvolvidas diversas formas de expansão da consulta baseadas na combinação destes conceitos.

Quando o descritor é formado por um conjunto de termos, a expressão agrega uma informação adicional, que é a seqüência de aparecimento dos termos na expressão. A fim de explorar esta nova evidência, os documentos e consultas são representados em um espaço vetorial multidimensional de termos, mas também de expressões definidas no tesauro. A fim de avaliarmos a similaridade entre os documentos e a consulta, a mesma representação espacial

será aplicada aos documentos. A partir dessa abordagem, as consultas serão sempre expandidas com descritores definidos no tesauro, mas nunca com os termos em separado que compõem estes descritores. Portanto, quando dissermos que a consulta foi expandida com o descritor USE, significa que a consulta será expandida apenas pela expressão que corresponde ao descritor posicionado após o termo USE.

Na seção anterior, vimos que as relações entre os descritores são definidas pelos operadores USE, UP, TG, TE e TR. A partir desta informação, podemos expandir a consulta de 8 formas diferentes. Experimentalmente, determinamos a melhor forma de expansão de consulta em nosso contexto.

Considerando que as consultas originais são compostas apenas por termos que, no conjunto, definem um conceito do tesauro, a primeira forma de expansão da consulta será adicionar à consulta este conceito.

No meio jurídico, devido à falta de padronização das expressões de categorização dos documentos e fatos jurídicos, a expansão da consulta com os sinônimos e equivalentes do conceito principal (indicados pelos operadores USE e UP), intuitivamente, produzirá um aumento da qualidade das respostas retornadas, o que se confirmou nos testes, conforme veremos mais adiante. Portanto, adicionalmente à expansão com o descritor, podemos utilizar também expansão com descritores associados, conforme indicado pelos descritores USE + UP. As outras 7 formas de expansão serão obtidas por combinação com outros operadores. Cada uma das combinações entre os descritores referentes aos operadores TG1, TE1 e TR produzirá uma forma diferente de expansão da primeira consulta.

4. A Jurisprudência

O documento jurídico, aqui referido como a jurisprudência, é um documento jurídico criado pelos órgãos judiciários brasileiros e que é de interesse público.

O documento é composto por três campos : *Ementa; Indexação; Acórdão*. A *Ementa* e o *Acórdão* são redigidos pelo juiz ou juizes no processo judicial de onde emanou a jurisprudência. A *Ementa* se constitui em um texto resumido do *Acórdão* que representa o tema central da jurisprudência. A *Indexação* é incluída posteriormente ao processo judicial, a fim de disponibilizar a jurisprudência em sistemas informatizados de recuperação de informação. Profissionais especializados em documentação, após lerem a *Ementa* e o *Acórdão*, selecionam os conceitos jurídicos do tesauro que fazem parte do tema central da jurisprudência e os relacionam no campo *Indexação*. O texto do *Acórdão*, por sua vez, inclui uma discussão jurídica do processo judicial, que envolve muitas teses não necessariamente relacionadas ao tema central do processo. Muitos termos fazem parte de citações doutrinárias, ou leis, apresentadas como argumentação, mas que não tem relação direta com o tema central da jurisprudência. Esta questão é importante para se perceber que cada um destes campos provê uma evidência semântica de natureza distinta.

5. Modelo Vetorial utilizado para cálculo do ranking

O modelo vetorial utilizado para calcular uma ordenação dos documentos com relação a uma consulta do usuário considera que documentos e consultas são indexados por termos. A cada termo k_i do documento d_j é atribuído um peso w_{ij} que é usualmente baseado no valor tf-idf (term frequency – inverse document frequency) definido em [7]. A cada termo ou descritor k_i da consulta também é atribuído um valor w_{iq}. A similaridade $Sim(d_j|q)$ do documento d_j em relação à consulta q pode ser computada pelo coseno do ângulo entre os dois vetores conforme a equação abaixo.

$$Sim(d_j|q) = \frac{\sum w_{ij}.w_{iq}}{\sqrt{\sum_{\forall i}(w_{ij})^2}\sqrt{\sum_{\forall i}(w_{iq})^2}} \quad \forall i$$

Podemos aplicar o modelo de espaço vetorial ao problema jurídico, o que implica que a estrutura do documento jurídico (ou seja, suas seções) é desconsiderada. Ademais, informação semântica do tesauro não é considerada.

6. Algoritmos de Combinação Bayesiana

A fim de melhorar a qualidade das respostas retornadas com a expansão da consulta, existe outra evidência a ser explorada : a semântica contida em cada um dos campos do documento jurídico.

Ao considerarmos a estrutura do documento jurídico, obtemos 4 formas distintas de ordenação, que são : *Documento-Plano* (considera todo o texto do documento como se não houvesse estrutura); *Documento-Ementa* (considera apenas o texto do campo *Ementa*); *Documento-Indexação* (considera apenas o texto do campo *Indexação*); *Documento-Acórdão* (considera apenas o texto do campo *Acórdão*).

As redes bayesianas nos permitem modelar o problema em um arcabouço conceitual que naturalmente representa todas estas evidências. Para o entendimento adequado da solução adotada, são abordados, nesta seção, dois modelos de redes bayesianas : o modelo de rede crenças (genérico) e o específico para o problema.

6.1 O Modelo de Rede de Crenças para Recuperação de Informação

Esta seção mostra como modelar uma solução baseada em conteúdo para o problema de recuperação de informação utilizando redes bayesianas. Para esta tarefa, adotaremos o modelo de rede de crenças definido em [8]. Este modelo possui uma visão epistemológica (em oposição à uma visão frequentista) do problema de recuperação de informação e interpreta as probabilidades como níveis de crença destituídos de experimentação, da mesma forma que em [9,10]. Esta é a razão para chamá-lo de modelo de rede de crenças. O modelo de rede de crenças adota os redes bayesianas como embasamento teórico. Redes bayesianas são úteis porque fornecem um formalismo gráfico para explicitar a representação das independências entre as variáveis da distribuição probabilística conjuntiva pertinentes ao problema de Recuperação de Informação, que são : os termos; os documentos; a consulta (ver figura 3).

Em um sistema tradicional de recuperação de informação baseado em conteúdo, os documentos e as consultas do usuário são normalmente representadas por um conjunto de termos. Como resultado desta interpretação, a consulta e os documentos são representados de forma análoga como proposto em [8]. A figura 3 ilustra uma rede de crenças que reflete a simetria. Nesta rede, cada nó Dj modela um documento Dj, o nó Q modela a consulta do usuário Q, e os nós k_i modelam os termos encontrados na coleção.

O vetor k é utilizado para referenciar qualquer um dos estados possíveis dos nós raiz ki, que são os nós sem pais. Uma variável aleatória binária está associada ao nó Q, também denotada por Q. Nesta notação, sempre está claro quando estamos nos referindo à consulta, ao nó na rede, ou ao valor binário da variável associada. A variável Q é 1, denotada por *q*, para indicar que Q está ativa e Q=0, denotada por $\bar{q}$, para indicar que a variável Q está inativa. Analogamente, uma variável aleatória D_j é associada com o documento nó D_j. A variável D_j é

1, denotada por d_j, para indicar que D_j está ativa e D_j=0, denotada por $\bar{d}_j$, para indicar que a variável está inativa. Uma variável aleatória binária k_i também está associada com cada termo k_i. Todas essas variáveis são binárias devido à necessidade de uma representação simples e também por fornecer semântica suficiente para modelar o problema de recuperação de informação. A variação dos níveis de relevância é representada no modelo como probabilidades condicionais, como discutiremos mais adiante.

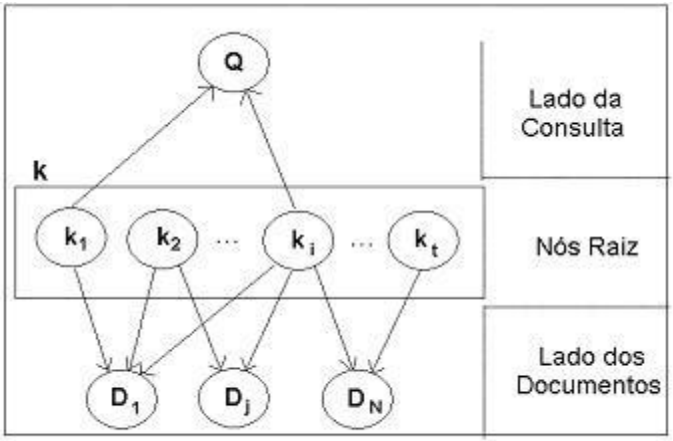


Figura 3 – Modelo de rede de crenças

A instanciação dos nós raiz ($\boldsymbol{k}$) separa os nós dos documentos (D) dos nós da consulta (Q), fazendo-os mutuamente independentes (veja a teoria bayesiana para mais detalhes em [10]). Então, na rede de crenças da figura 3, dizemos que a consulta está no lado da consulta da rede, enquanto os documentos estão no lado dos documentos da rede.

Na rede da figura 3, o cálculo do *ranking* é baseado na quantificação da similaridade entre um documento D_j e a consulta Q dado pela probabilidade P(D_j=1|Q=1), ou simplesmente por P(d_j|q) (probabilidade de que a variável aleatória D_j esteja ativa dado que a variável aleatória Q está ativa). Pela regra das probabilidades totais e as independências modeladas na rede, podemos escrever :

$$P(d_j \mid q) = \bullet \sum_{K} P(d_j \mid \boldsymbol{k}) \, P(q \mid \boldsymbol{k}) \, P(\boldsymbol{k}) \qquad (1)$$

onde • é uma constante de normalização [7]. Esta é uma expressão genérica para cálculo do *ranking* de um documento D_j com relação à consulta Q, em nosso modelo de rede de crenças.

Modelo de Espaço Vetorial em uma Rede de Crenças

Para computar um *ranking* vetorial na rede de crenças, adotamos uma especificação particular para as probabilidades a priori de P(***k***), P(q | ***k***) e P(d_j | ***k***). A probabilidade a priori P(***k***) será calculada pela equação abaixo :

$$P(k) = \begin{cases} 1 \text{ se } \forall i \; g_i(q) = g_i(k) \\ 0 \text{ senão} \end{cases} \qquad (2)$$

onde g_i(u) é uma função que retorna o estado (0 ou 1) da i-ésima variável no vetor u. Equação (2) estabelece que apenas os termos/expressões contidos na consulta Q serão levados em conta para cálculo do modelo vetorial.

Para a probabilidade P(q|***k***) escrevemos :

$$\begin{cases} 1 \text{ se } \forall_i g_i(q) = g_i(k) \end{cases} \qquad (3)$$

$$P(q \mid k) = \begin{cases} \square \\ \square\square\ 0 \quad \text{sen } \tilde{a}o \end{cases}$$

Para a probabilidade P($d_j \mid \boldsymbol{k}$) escrevemos :

$$P(d_j \mid k) = \frac{\sum_{i=1}^{t} w_{ij} . w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad (4)$$

onde w_{ij} e w_{iq} são os pesos tf-idf [12,13] usados no modelo vetorial. Pela substituição das equações (2) a (4) em (1), obtemos o *ranking* para os documentos D_j, expressos por P($d_j \mid q$), que preservam a mesma ordem ditada pelo *ranking* vetorial.

6.2 Adaptando o Modelo Bayesiano para o Problema Jurídico

Por meio da rede bayesiana do item anterior, são modeladas algumas evidências específicas do problema jurídico para, a partir daí, agregar novas evidências, o que produzirá um novo modelo. Pelo fato do modelo expandido representar uma quantidade maior de evidências pertinentes ao problema jurídico, pretende-se obter uma melhor qualidade das respostas retornadas.

O modelo genérico de redes bayesianas modela o problema jurídico nos seguintes aspectos : consulta(*q*), termos e expressões(**k**), e *Documento-Plano*(*D*). Expandiremos a rede bayesiana discutida no item anterior a fim de representar as evidências referentes aos campos do documento (*Ementa, Indexação, Acórdão*). Isto é realizado pela adição de novas arestas, nós e probabilidades à rede bayesiana apresentada na figura 3 . Esta expansão é modular no sentido de que preserva todas as propriedades da rede anterior e ainda assim incorpora as novas evidências contidas em cada um dos três campos do documento.

De acordo com a figura 4, que incorpora as novas evidências, o lado esquerdo representa o modelo anterior, cujas variáveis aleatórias sofreram as seguintes adaptações : a variável k_i representa termos genéricos e expressões associadas a conceitos jurídicos definidos como descritores no tesauro; a variável D_j foi renomeada para DP_j. Esta evidência será referenciada ao longo deste trabalho como *Documento-Plano* (O nome caracteriza o fato da variável aleatória DP_j estar associada ao nó DP_j, que representa a contribuição do conteúdo do documento completo, sem informações sobre a estrutura, para a geração do *ranking*) .

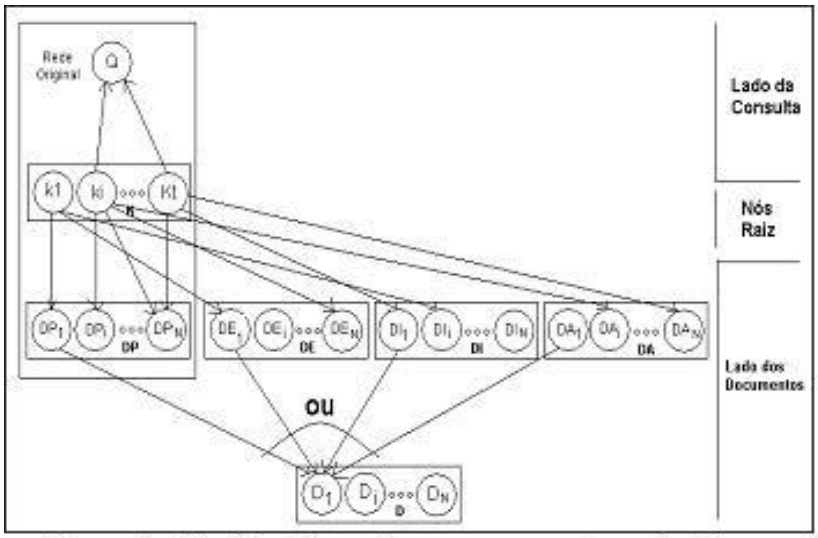


Figura 4 - Modelo Bayesiano para o problema jurídico

O novo modelo também nos permite

modelar a expansão da consulta, conforme descrito na seção 3 : as arestas que conectam o nó da consulta *Q* ao grupo de nós ***k*** modelam a evidência de que a consulta é formada não apenas por termos, mas também por conceitos e características associadas a estes conceitos (definidos conforme sintaxe de construção do tesauro jurídico). A instanciação de uma determinada consulta *Q* se dá pela instanciação de cada um de seus termos, mas também pelo descritor do tesauro contido em k_i que é formado pelo conjunto de termos da consulta. Além dessa expansão, as respectivas características deste descritor também serão utilizadas na expansão (sinônimos, termos genéricos, termos específicos e termos relacionados).

Com o objetivo de modelarmos os campos dos documentos, foram criados 4 novos grupos de nós, contidos no lado direito da rede bayesiana da figura 4:

- grupo de nós DE que modelam a evidência denominada *Documento-Ementa*, onde cada variável associada DE_j representa o texto do campo *Ementa* de um determinado documento D_j;
- grupo de nós DI que modelam a evidência denominada *Documento-Indexação*, onde cada variável associada DI_j representa o texto do campo *Indexação* de um determinado documento D_j;
- grupo de nós DA que modelam a evidência denominada *Documento-Acórdão*, onde cada variável associada DA_j representa o texto do campo *Acórdão* de um determinado documento D_j;
- grupo de nós *D*, combinam as 4 fontes de evidência para gerar uma combinação final para o documento D_j .

As arestas que apontam do conjunto ***k*** para os nós DE, DI, DA e DP permitem associar os termos e expansões da consulta a cada uma das representações do documento D_j. Da mesma forma, as arestas dos nós DE, DI, DA e DP para os nós D nos permitem associar estas evidências a cada um dos documentos D_j.

Cada variável aleatória DE_i, DI_i e DA_i, associadas aos respectivos nós DE_i, DI_i e DA_i, representa o documento como se este fosse composto exclusivamente por um dos campos da jurisprudência. A variável aleatória DE_1, por exemplo, corresponde apenas ao texto da *Ementa* do documento número 1, a variável aleatória DI_2 se refere ao texto da *Indexação* do documento número 2, e a variável aleatória DA_7 representa apenas o texto do *Acórdão* do documento número 7.

O conjunto de nós ***k*** é usado para modelar a ocorrência dos termos e conceitos associados à consulta *Q* e, uma vez instanciados, induzem à crença em cada um dos nós nos conjuntos DP, DE, DI, e DA. A propagação destas crenças é feita de acordo com as probabilidades condicionais regidas pelo relacionamento entre o conjunto ***k*** e cada um dos conjuntos DP, DE, DI e DA. Estas probabilidades condicionais são especificadas pelo modelo vetorial anteriormente discutido.

Seguindo a natureza conceitual e probabilística do grafo de Bayes, faremos a seguinte consideração : A cada nó DE_j de DE está associada, respectivamente, uma variável binária aleatória DE_j. Esta variável é 1 (*Dej*) para indicar que a evidência do campo *Ementa* associada ao documento D_j será considerada para o processamento do *ranking*. Esta evidência não será considerada para a geração do *ranking* quando o seu valor for 0 (zero), indicada por $\neg De_j$. Da mesma forma, a cada nó DI_j de DI está associada, respectivamente, uma variável binária aleatória DIj. Esta variável é 1(Di_j) para indicar que a evidência do campo *Indexação* associada ao documento D_j será considerada para o processamento do *ranking* . Analogamente, a cada nó DA_j de DA está associada, respectivamente, uma variável binária aleatória DAj. A variável é 1

(*Daj*) para indicar que a evidência do campo *Acórdão* associada ao documento D_j será considerado para o processamento do *ranking*. Da mesma forma, a cada nó DP_j de DP está associada, respectivamente, uma variável binária aleatória DP_j. Esta variável é 1 (*Dpj*) para indicar que a evidência do *Documento-Plano* associada ao documento D_j será considerada para o processamento do *ranking*.

A manipulação de algumas probabilidades, conforme será visto mais adiante, permite que consideremos, ou não, cada um dos grupos de variáveis aleatórias DP, DE, DI e DA. Dessa forma, podemos criar 15 modelos diferentes, onde as evidências poderão ser comparadas de modo separado ou combinado para a geração de um *ranking* final.

6.3 Equação geral para cálculo do ranking

Na figura 4, o *ranking* $P(d_j|q)$ associado ao documento D_j pode ser computado utilizando a equação (1). Entretanto, a probabilidade condicional $P(d_j|\boldsymbol{k})$ depende da operação disjuntiva entre as evidências *Documento-Plano*, *Documento-Ementa*, *Documento-Indexação* e *Documento-Acórdão*, conforme mostrado no modelo. Isto é realizado da seguinte forma :

$$P(dj \mid \boldsymbol{k}) = P(DP \vee DE \vee DI \vee DA \mid \boldsymbol{k}) \quad (5)$$

$$P(d_j \mid \boldsymbol{k}) = 1 - (1 - P(Dpj \mid \boldsymbol{k})) \times (1 - P(Dej \mid \boldsymbol{k})) \times (1 - P(Dij \mid \boldsymbol{k})) \times (1 - P(Daj \mid \boldsymbol{k})) \quad (6)$$

Substituindo a equação (6) na (1), escrevemos :

$$P(dj \mid \boldsymbol{k}) = \eta \sum [1 - (1 - P(Dpj \mid \boldsymbol{k})) \times (1 - P(Dej \mid \boldsymbol{k})) \times (1 - P(Dij \mid \boldsymbol{k})) \times (1 - P(Daj \mid \boldsymbol{k}))] \times P(q \mid \boldsymbol{k}) \times P(\boldsymbol{k}) \quad (7)$$

O cálculo das probabilidades $P(d_j \mid q)$ depende dos estados das variáveis DP_j, DE_j, DI_j, e DA_j.

A probabilidade $P(q|\boldsymbol{k})$ pode ser computada utilizando os estados dos nós raiz k_i. Por meio da especificação dos estados de todos estes nós, podemos estabelecer alternativas interessantes para cálculo do *ranking* do documento D_j em relação à consulta Q.

6.4 Alternativas de combinação dos rankings

A 15 combinações entre as quatro evidências DP, DE, DI e DA são conceitualmente descritas nesta seção e avaliadas em nossos experimentos, onde, empiricamente, é determinada a combinação que produz o resultado de melhor qualidade.

Como discutido na Seção 6.1, o modelo de redes de crença pode representar o modelo vetorial por meio da especificação de probabilidades condicionais na rede. Para simplificar a nossa anotação, seja R_{jq} uma referência para o *ranking* vetorial do documento D_j com relação à consulta *Q* computada de acordo com o nosso modelo de rede utilizando a equação (4). Então,

$$R_{jq} = \frac{\sum_{i=1}^{t} w_{ij} . w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (8)$$

A fim de evitar a exaustiva repetição da conceitualização das 15 ordenações possíveis em nossa rede Bayesiana, é apresentada apenas uma opção, donde pode-se, facilmente, concluir como serão criadas as 14 restantes. Cada uma das opções de ordenação das respostas é criada pela alternância de consideração de cada uma das quatro evidências *Documento-Plano*; *Documento-Ementa*; *Documento-Indexação* e *Documento-Acórdão*. Isto se faz pela manipulação das probabilidades P(*Dpj*|***k***), P(*De*$_j$ |***k***), P(*Dij*|***k***) e P(*Daj*|***k***). Tornando qualquer uma destas probabilidades igual a zero, em (7), desconsidera-se a contribuição da respectiva evidência para geração do *ranking*.

Como exemplo, para representar o *ranking* baseado apenas no conteúdo do documento plano, evidência *Documento-Plano*, deve-se desconsiderar a contribuição das demais evidências (*Documento-Ementa*, *Documento-Indexação* e *Documento-Acórdão*). Isto se faz pela definição das seguintes probabilidades :

$$P(De_j \mid \boldsymbol{k}) = 0;\ P(Di_j \mid \boldsymbol{k}) = 0;\ P(Da_j \mid \boldsymbol{k}) = 0 \qquad (9)$$

Seja R_{jq} da equação (8) o *ranking* computado pelo modelo vetorial onde se considera todo o texto do documento plano, obtido pela probabilidade P(*Dp*$_j$ | ***k***), e aplicando as equações (2), (3), (8), (9), em (7), obteremos :

$$P(d_j \mid q) = \bullet \,.\, R_{jq} \qquad (10)$$

Portanto, a rede geral da figura 4 naturalmente representa o *ranking* implementado pelo modelo de espaço vetorial.

A partir do exemplo acima, pode-se concluir o modo de desenvolvimento das quatorze opções restantes do total das quinze possíveis.

7. Resultados experimentais

Nesta seção, é apresentado o ambiente de testes e são analisados os resultados obtidos com o protótipo desenvolvido. Também comparamos graficamente os resultados obtidos pelo modelo vetorial de aplicação genérica, com as diversas opções de expansão da consulta e também as 15 possibilidades diferentes de ordenação das respostas em nosso arcabouço bayesiano, apresentadas anteriormente.

7.1 A coleção de documentos, as consultas e características de implementação

Para efetuarmos os testes, utilizamos uma coleção de 155.000 documentos composta por jurisprudências do Supremo Tribunal Federal e Tribunal de Alçada de Minas Gerais.

A seção *'10-Apêndice'* contém a lista das 25 consultas utilizadas nos testes. Para cada uma das curvas apresentadas nos gráficos a seguir, fez-se uma avaliação manual de relevância dos 50 primeiros documentos retornados por cada uma das nossas 43 possibilidades de ordenação de respostas (4-Modelo Vetorial Tradicional + 28-expansão da consulta + 11combinações bayesianas) por consulta. Assim, para cada consulta, avaliamos um mínimo de 50 documentos e um máximo de 2150 documentos na resposta. A partir daí, obteve-se a média dos valores de precisão das consultas para cada um dos 10 valores de revocação.

A implementação possui as seguintes características : eliminação de StopWords (termos sem conteúdo semântico, tais como, preposições e artigos) e, no parsing(algoritmo de varredura de termos e expressões sobre a coleção de documentos), foram criadas listas do tipo (Documento, Freqüência) também para as expressões que representam descritores do tesauro do CJF, conforme descrito anteriormente.

Pelo fato do tesauro padronizar a utilização dos seus termos no modo singular, implementou-se uma rotina de conversão, onde o termo convertido para o singular é considerado como existente no documento se também for encontrado no tesauro do CJF.

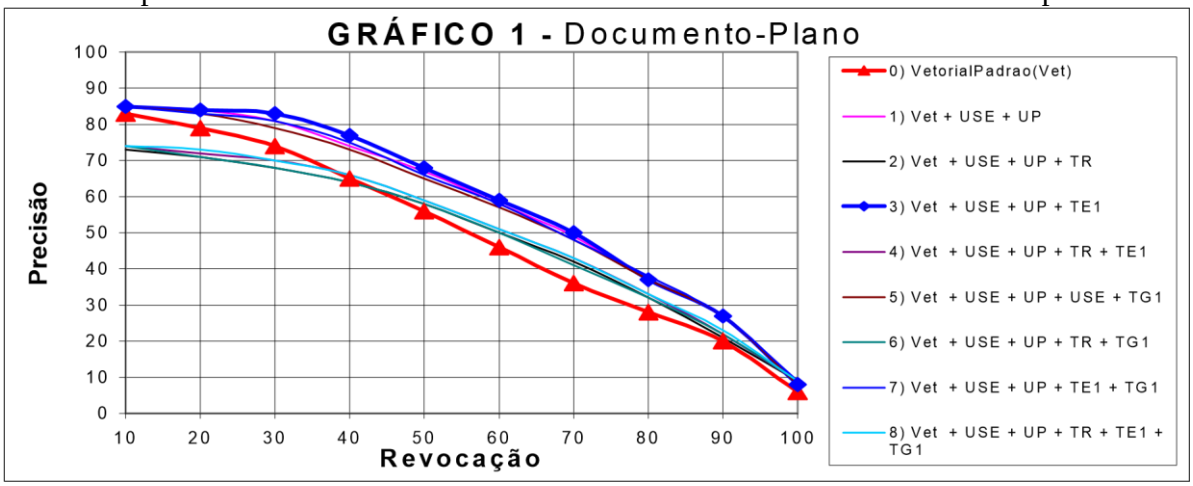
7.2 Resultados Obtidos

As duas técnicas de recuperação de informação (RI) utilizadas neste trabalho, expansão da consulta e combinação dos resultados, são complementares, pois, os resultados combinados são os que se utilizam da expansão da consulta. Dessa forma, primeiramente, determinamos a melhor forma de expansão da consulta para, em seguida, combinarmos os resultados. A obtenção dos resultados em cada uma das duas etapas nos permite verificar que a combinação bayesiana foi capaz de melhorar os resultados obtidos com a expansão da consulta.

A mensuração da qualidade das respostas retornadas é feita por meio de dois índices : precisão e revocação, que são duas medidas quantitativas, utilizadas também para comparar as respostas geradas por um algoritmo automático com as respostas indicadas por especialistas na área em questão. (que são chamados *"documentos relevantes"*). Precisão é uma medida da fração de documentos relevantes (ou seja, dos documentos indicados pelos especialistas) que foram recuperados pelo algoritmo sendo avaliado. Maiores detalhes, obtêm-se em [12].

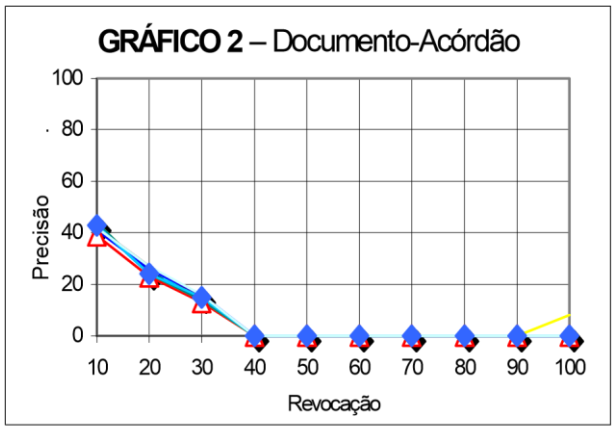
7.2.1 Expansão da consulta

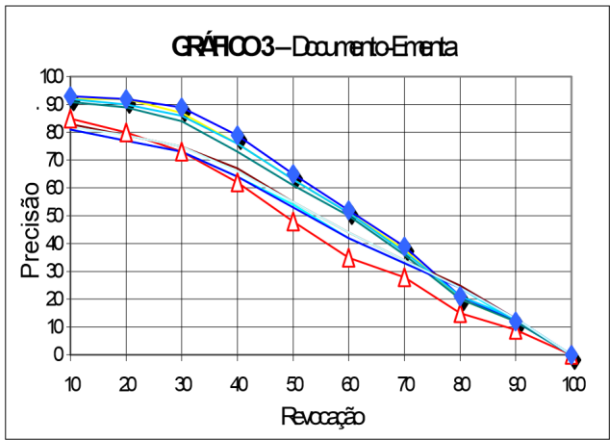
Os gráficos de 1 a 4 abaixo mostram os resultados obtidos pela expansão da consulta com cada um dos 4 tipos de documentos. Estes se referem às evidências relativas aos 3 campos da



jurisprudência, mais o documento completo, respectivamente : *Documento-Acórdão*, *Documento-Ementa*, *Documento-Indexação e Documento-Plano* (ver seção 6.2).

Em todos os 4 gráficos, a curva de número zero representa o modelo vetorial padrão (técnica genérica de RI) e as outras 8 curvas representam as diferentes formas de expansão da consulta, descritas na seção 3.



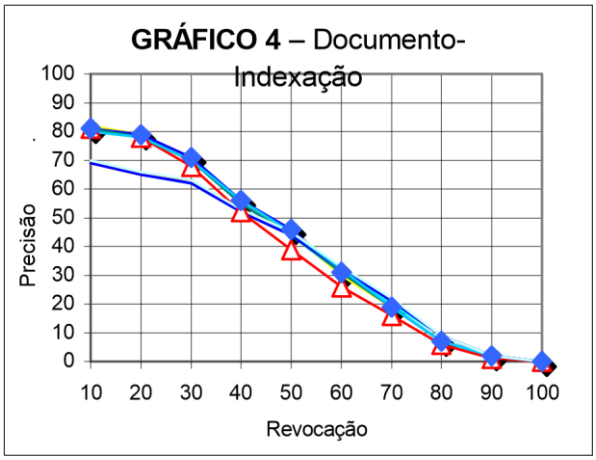


De acordo com a legenda do *GRÁFICO 1*, válida para todos os 4 gráficos, a entrada *'8)Vet + USE + UP + TR + TE1+TG1'*, por exemplo, designa a curva de número 8, e especifica que a consulta original, composta por termos, foi expandida pelos seguintes descritores : descritor que representa a consulta (expansão comum às consultas de número 1 a 8); descritores sinônimos e equivalentes, tipos USE e UP (expansão comum às consultas de número 1 a 8); descritores relacionados (TR), específicos(TE1) e genéricos(TG1) . Seguindo o mesmo raciocínio, obtêm-se a forma de expansão das demais consultas.

De acordo com os gráficos 1, 3 e 4, para todos os valores de revocação, existem diversas formas de expansão da consulta cuja qualidade das respostas retornadas é superior ao modelo vetorial padrão.

Outra verificação se refere ao gráfico número 2 : O Acórdão do documento, apesar de conter muito conteúdo informativo, produziu, de um modo geral, baixa precisão, igual a zero para valores de revocação iguais ou superiores a 40%.

Ao compararmos o gráfico número 4 aos de números 1 e 3 verificamos que o *ranking Documento- Indexação*, por si só, produziu uma qualidade das respostas retornadas inferior em relação aos *rankings Documento-Plano* e *Documento-Ementa*. Isto nos indica que a qualidade da categorização manual pode ser melhorada por meio de técnicas de recuperação de informação.



As melhores curvas de expansão da consulta, de acordo com os gráficos, são as curvas de número 3 dos gráficos 1 e 3. A curva 3 do gráfico 1 produziu o melhor resultado para os valores de revocação maiores ou iguais a 50 %, enquanto a do gráfico 3 produziu melhores resultados para os valores de revocação menores ou iguais a 40 %.

Propositadamente, realçamos as curvas de número 0(zero) e a de número 3, em todos os gráficos, a fim de que se perceba, mais claramente, que a curva *'Vet + USE + UP + TE1'* (ou seja, a entrada 3) representa a forma de expansão da consulta cuja qualidade das respostas retornadas produziu a melhor

resultado nos gráficos 1, 3 e 4. O gráfico número 2 não foi considerado porque as suas curvas praticamente se coincidem, portanto não existe uma definição clara do melhor resultado.

Para avaliarmos numericamente os valores de cada curva em relação à curva do modelo vetorial padrão (curva de número 0, técnica genérica de RI), escolheu-se os dados do gráfico 1, *Documento-Plano*, que apresenta os melhores valores de precisão, relacionados na tabela 1 abaixo. A primeira linha superior da tabela contém o número das curvas descritas na legenda do gráfico 1. As colunas de título 'Pr' contêm os valores de precisão e as linhas contêm os valores de revocação (10 a 100).

A coluna cujo título é o símbolo '• ' contém a diferença entre os valores de precisão da curva atual e a curva de número 0(zero) tomada como referência. A linha de título 'Soma(Pr)/10' representa a média dos valores de precisão de cada curva.

Curva	0	1		2		3		4		5		6		7		8	
Revocação	Pr	Pr	Δ	Pr	Δ	Pr	Δ	Pr	Δ	Pr	Δ	Pr	Δ	Pr	Δ	Pr	Δ
10	83	85	2	73	-10	85	2	74	-9	85	2	74	-9	85	2	74	-9
20	79	84	5	71	-8	84	5	72	-7	83	4	71	-8	83	4	73	-6
30	74	81	7	68	-6	83	9	70	-4	79	5	68	-6	81	7	70	-4
40	65	74	9	64	-1	77	12	66	1	73	8	64	-1	75	10	66	1
50	56	67	11	58	2	68	12	59	3	65	9	58	2	66	10	59	3
60	46	58	12	50	4	59	13	51	5	57	11	50	4	58	12	51	5
70	36	49	13	42	6	50	14	43	7	48	12	41	5	48	12	43	7
80	28	37	9	32	4	37	9	33	5	37	9	32	4	38	10	33	5
90	20	27	7	21	1	27	7	22	2	27	7	22	2	27	7	23	3
100	6	8	2	9	3	8	2	9	3	8	2	9	3	9	3	9	3
Soma(Pr)/10	49,30	57,00		48,80		**57,80**		49,90		56,20		48,90		57,00		50,10	

Tabela 1 - Dados do gráfico 1

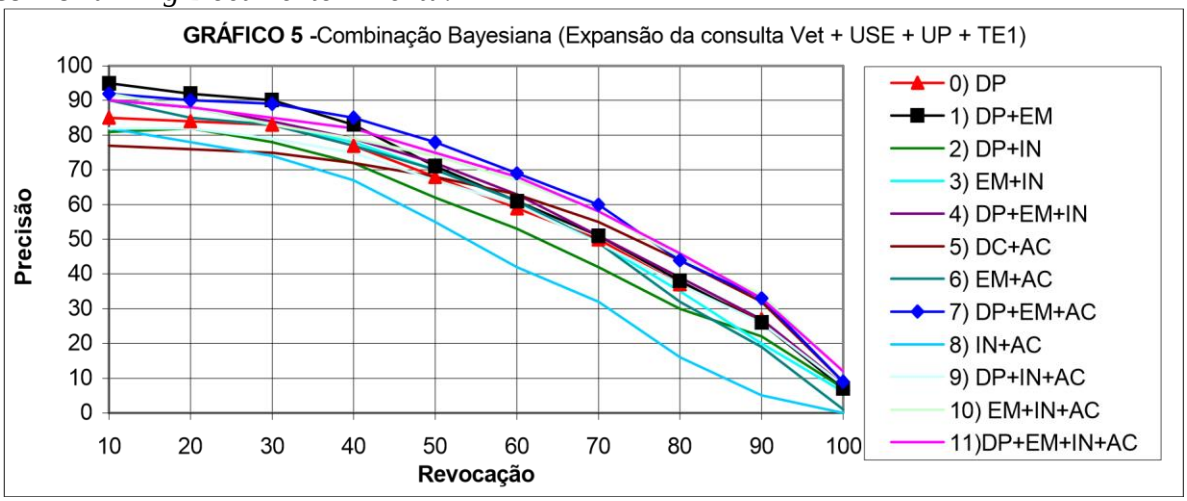
Ao analisarmos a tabela 1, podemos fazer as seguintes considerações :

- a expansão da consulta com os sinônimos e equivalentes (USE + UP) , curva número 1, produziu resultados iguais ou superiores ao vetorial padrão para todos os valores de revocação. O sucesso dessa estratégia mostra que o conteúdo informacional do tesauro jurídico do CJF, no que se refere aos sinônimos e equivalentes, é importante para a recuperação de documentos jurídicos;
- as curvas de número 2, 4, 6 e 8, para valores de revocação menores do que 40%, produziram valores de precisão inferiores ao vetorial padrão, enquanto as curvas de número 1, 3, 5 e 7 sempre produziram valores maiores ou iguais aos do vetorial padrão. A partir destes resultados podemos afirmar que a expansão da consulta com os descritores do tipo TR tendem a piorar a qualidade das respostas retornadas, pois, em todas as curvas onde as consultas foram expandidas com esta definição, ocorreram valores de precisão inferiores aos do vetorial padrão;
- de acordo com os valores da linha *'Soma(Pr)/10'*, somada à análise gráfica das curvas nos gráficos 1, 3 e 4, podemos concluir que a curva de número 3 apresentou a melhor forma de expansão da consulta (formada pelos descritores do tipo USE, UP e TE1).

7.2.2 Combinação de rankings

Após a determinação, na seção anterior, de que a melhor forma de expansão da consulta é a que inclui os descritores do tipo USE, UP e TE1, combinamos os diversos *rankings* descritos na seção 6.2. Desconsideramos as alternativas IN, EM, e AC porque, quando aplicadas em separado, produzem resultados inferiores a DP.

O gráfico 5 abaixo apresenta 11 curvas de combinação dos *rankings*. Os *rankings* que são combinados para obtenção de cada curva são descritos na legenda e se utilizam das seguintes abreviações : *DP*, *ranking Documento-Plano*; *EM*, *ranking Documento-Ementa*, *IN*, *ranking Documento-Indexação*; *AC*, ranking *Documento-Acórdão*. O literal *'1) DP+EM'*, por exemplo, designa que a curva de número 1 é composta pela combinação do *ranking Documento-Plano* com o *ranking Documento-Ementa*.



A curva de número 0(zero) do gráfico 5 representa o melhor resultado obtido com a expansão da consulta : curva 3 do gráfico 1 (*Documento-Plano*). Esta curva nos permitirá visualizar, graficamente, que os resultados anteriormente obtidos foram melhorados.

Ao analisarmos o gráfico 5, existem duas curvas que se destacaram com os melhores resultados a de número 1 e a de número 7. A curva 1 possui maior precisão para os valores de revocação menores ou iguais a 30%, enquanto a curva 7 é melhor para todos os valores entre 35-75% de revocação. Além disso, é importante verificar que a curva 7 possui maior precisão do que a curva 0(zero) para todos os valores de revocação. A partir desta análise, podemos concluir que a curva *'7)DP+EM+AC'* produziu os melhores resultados.

A inclusão do *ranking Documento-Ementa*(EM) na composição do melhor *ranking* combinado representa a maior importância do campo Ementa em relação ao Acórdão e à Indexação para relevância do documento. Da mesma forma, a inclusão do *ranking Documento-Acórdão*(AC) na composição do melhor *ranking* combinado mostra que o grande conteúdo informativo do campo Acórdão contribui, de forma contundente, para a qualidade da recuperação de informação contida em Jurisprudências.

8 – Conclusão

Neste trabalho, propusemos e avaliamos um modelo de recuperação de documentos jurídicos baseado na combinação de evidências obtidas a partir do texto do documento, de sua estrutura e de um tesauro jurídico elaborado pelo Conselho de Justiça Federal (CJF). O modelo é representado em um arcabouço de Redes Bayesianas de Crenças, porque este arcabouço se revelou de grande utilidade em problemas de natureza similar mas que ocorrem em outros contextos.

Para a avaliação do modelo, utilizamos uma coleção composta por 155.000 documentos jurídicos e 25 consultas de referência, selecionadas por nós. Quinze(15) formas distintas de recuperar (e ordenar) as respostas, todas elas geradas dentro de nosso modelo de Redes Bayesianas de Crenças, foram avaliadas e comparadas. A tabela 2 apresenta um sumário de nossas conclusões, como se segue.

Revocação	10	20	30	40	50	60	70	80	90	100
Precisão - VetorialPadrão(Vet)	83	79	74	65	56	46	36	28	20	6
Precisão - Melhor resultado	92	90	89	85	78	69	60	44	33	9
Ganho Percentual	10,84	13,92	20,27	30,77	39,29	50,00	66,67	57,14	65,00	50,00

Tabela 2 - Comparativo entre os resultados obtidos pelo modelo vetorial e pelo modelo Bayesiano que combina as evidências sobre o *Documento Plano* (DP), *Ementa* (EM) e *Acórdão* (AC) (rotulado *`Melhor Resultado`*)

Na tabela 2, a linha rotulada *`Vetorial Padrão`* apresenta os resultados de precisão média (para as 25 consultas) gerados pela aplicação do modelo vetorial clássico ao problema de recuperar jurisprudências. A linha rotulada *`Melhor Resultado`* apresenta os resultados de precisão média (para as 25 consultas de referência) gerados pelo melhor de nossos algoritmos Bayesianos (combinar evidências do documento *Plano*(DP), da *Ementa*(EM) e do *Acórdão*(AC)). A linha rotulada *`Gasto Percentual`* indica o ganho de precisão (para cada faixa padrão de revocação) obtida por nosso modelo Bayesiano baseado em múltiplas evidências. Observamos que o ganho de precisão variou cerca de 10% a mais de 65%.

Este trabalho mostrou, portanto, que o algoritmo especializado, sob o ponto de vista da qualidade das respostas retornadas, produz resultados muito melhores do que o genérico (modelo vetorial tradicional) porque é capaz de agregar evidências inerentes ao problema específico. Sob o ponto de vista prático, a técnica desenvolvida mostra potencialmente valiosa para a implementação de sistemas aplicativos especializados na área jurídica. Ademais, este trabalho mostrou não apenas que o tesauro do CJF possui importante conteúdo informacional a ser explorado em sistemas jurídicos, mas também apresentou uma técnica eficaz de explorálo.

9. Referências

[01] Aurélio Buarque de Holanda. Novo Dicionário da Língua Portuguesa. 2ª Edição, Editora Fronteira, 1986.

[02] Sharon L. Greene, Susan J. Devlin, Philip E. Cannata, and Louis M. Gomez. No Ifs, ANDs, or Ors : A study of database querying. *International Journal of Man-Machine Studies*, 32(3):303-326, 1990.

[03] http:\\www.cjf.gov.br.

[04] A. Pollock and A. Hockley. What's wrong with Internet searching. *D-Lib Magazine*, March 1997.

[05] B. J. Jensen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study ou user queries on the Web. *ACM SIGIR Forum*, 32(1):5-17, 1998.

[06] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large alta vista query log. Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, CA, USA, 1998.

[07] Pearl, J., Probabilistic Reasoning in Intelligent Systems : Networks of PlausibleInference. Morgan Kaufman, San Mateo, CA, 1988.

[08] B. Ribeiro-Neto, I. Silva, and R. Muntz. Bayesian network models for IR. In *Proc. Of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253-260, Zurich, Switzerland, 1996.

[09] H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187-222,1991.

[10] S. Wong and Y. Yao. A probabilistic inference model for information retrieval. Information Sustems, 16(3): 301-321, 1991.

[11] B. Ribeiro-Neto, I. Silva, and R. Muntz. Bayesian network models for IR. *In Soft Computing in Information Retrieval : Techniques and Applications*. F. Crestani and G. Pasi editors, Springer Verlag. 2000.

[12] R. Baeza-Yates and B. Ribeiro Neto. *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999.

[13] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.

[14] I. Witten, A.Moffat, and T. Bell. *Managing Gigabytes : Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1999.

10. Apêndice - Consultas

A lista abaixo relaciona os 25 consultas utilizadas em nossos testes, cuja definição no tesauro do CJF pode ser encontrada em [3].

01) ABALROAMENTO
02) ABASTECIMENTO
03) ACIDENTE DE TRÂNSITO
04) BEM IMÓVEL
05) BACEN
06) CONTA-CORRENTE
07) ENRIQUECIMENTO ILÍCITO
08) HERANÇA
09) INSTITUIÇÃO FINANCEIRA
10) INVALIDEZ PERMANENTE
11) JUROS COMPOSTOS
12) MINISTÉRIO PÚBLICO
13) OBRIGAÇÃO SOLIDÁRIA
14) PENSÃO ALIMENTÍCIA
15) RESSARCIMENTO DO DANO
16) TELEFONE
17) TÍTULO DA DÍVIDA AGRÁRIA
18) VÍNCULO EMPREGATÍCIO
19) PLANO REAL
20) FOTOGRAFIA
21) PRESIDENTE DA REPÚBLICA
22) INSS
23) REINTEGRAÇÃO DE POSSE
24) DIREITO DO CONSUMIDOR
25)JUSTIÇA GRATUITA