Sensors & Transducers, Vol. 159, Issue 11, November 2013, pp. 255-262

Sensors & Transducers

http://www.sensorsportal.com

XML Data Retrieval Model Based On Two-dimensional Table Datasets

Lichuan Gu, Qingyan Guo, Youhua Zhang
School of Information & Computer, Anhui Agricultural University, Hefei 230036 Anhui, China
Tel.: 0551-65786188, fax: 0551-65786188
E-mail: lcgu@sohu.com

Received: 1 September 2013 /Accepted: 25 October 2013 /Published: 30 November 2013

Abstract: Retrieval problems of XML-based representation of data have been researched in this paper. In order to solve the large time and space overhead problem in building content index, this paper establish a data retrieval model advantageous to xml representation using the system automatically build two-dimensional table datasets. Take crop diseases and insect pests data for an example, this paper first gives the architecture of retrieval system based on XML crop diseases and insect pests' data; it also discusses about how to construct the two-dimensional table dataset and achieve the retrieval process; then it describes the text segmentation technique and the XSL style sheet conversion technology. Finally, under the VS.NET platform, using MVC design pattern develop and implement a prototype. *Copyright © 2013 IFSA.*

Keywords: Retrieval model, XML, Chinese word segmentation, Two-dimensional table datasets, Crop diseases and insect pests.

1. Introduction

Extensible Markup Language (XML) is a metadata markup language based on the Internet which is developed in recent years. It has advantages such as openness, platform-independent, self-descriptive, flexible, scalable and highly hierarchical data organization and form, now it is gradually replacing HTML, become the main form of information that resides on the Web. The benefits of using XML to represent data include: effective control and use of the information content of the web page; strengthen searching performance of data; enhance the quality of the data; strengthen the integration and interoperability between heterogeneous information systems and data sources; more flexible way to take advantage of the different ways to present the same data sources [1].

Because of this, the exploration, study and application of the XML from domestic and abroad during these years have been increasing, especially concerning on XML data retrieval research, [2-5]. XML data retrieval methods [6], include two methods. One is XML IR / query mode, use the query language for XML data like: Query, XPath to conduct XML data retrieval; another method is XML IR / keyword, extend the keyword search method of traditional text documents to the XML data, also can use the unique XML structure which is characterized as the auxiliary of retrieving information to limit the scope of the keywords of the node where. The former is mainly for data-centric XML data, the latter concentrating on text-centric XML data. To improve the efficiency of data retrieval for XML text, recently it is used mostly by building efficient indexing mechanism to achieve, including the contents of the index and structural index, efficient indexing mechanism is one of the main methods to achieve fast retrieval. However, building indexing, building content indexing, it needs a lot of time and space

Content indexing documents size is typically as 3 to 5 times large as that of the original XML document . Unlike traditional text documents, XML data search includes not only the content information, but also the structural information. In this paper, considering that the structure of the XML text data has a clear hierarchy and clear semantic features, a two-dimensional table text data retrieval model is proposed based on the XML datasets automatically building, which avoids the creation of content indexing, while the system build the two-dimensional table data set stored XML document data in a heap cache automatically, solve the overhead problem of time and space during the construction of content index. In this paper, the retrieval model is applied to crop pests' data retrieval, on the VS.NET platform, using the MVC design pattern; we developed an XML-based crop pest's data retrieval system. System shows that this method can quickly locate retrieve information to achieve the purpose of efficient retrieval.

2. XML Description of the Crop Pests Data

Currently, the data of Crop diseases, insect pests and weeds, are stored in a variety of heterogeneous databases, due to incompatibilities between different database, and most of the information standards are not uniform, which bring great difficulties to pests data retrieval and data migration, and reduce the efficiency of the information sharing. XML can easily integrate data from a variety of data sources, and data has good portability. In addition, crop pests data structure degree is not high, using the semi-structured data format standard XML to describe is undoubtedly a very good solution.

Take crops of rice for example, by looking though some of the information on rice pests, make analysis that the rice disease generally include the Chinese name, English name, common name, profile, symptoms, pathogens, control methods, the incidence of the conditions, the route of transmission, distribution, disease-resistant varieties, and several other part. This paper need to design these several parts as the XML element tags. The structure of an XML document is a typical layered structure, can be simulated as a tree, known as the structure tree. The tree contains a root node and a number of element nodes, element nodes both can express elements and attributes [7]. As shown in Fig. 1 is the tree of the rice disease, by the same method we can get structural tree of rice pests, weeds tree.

XML Schema is a specification used to describe the structure of XML documents. Marker of the rice pests in XML document is self definite, in order to ensure the smooth progress of the exchanging and sharing of data, the data structure of the XML tag and set must pass through XML Schema validation. According to the above-described structure, the structural diagram of rice diseases was transformed, definited .After the process, the XML Schema fragment was shown in Fig. 2.

The following is a specific XML document data fragment (rice disease section) This paper use data of crop pests, rice, peanuts, soybeans, wheat, common diseases, pests and weeds as example.

```
<Pest weeds Data>
<Disease ID = "D0001">
<Name>
    Chinese name "稻瘟病 </ Chinese name>
    <Name> Rice blast </ name>
    <Common name> blast, fire blast </ common name>
</ Name>
 Major damage <Introduction> leaf, stem, panicle ... </ Description>
 <Symptoms>  seedlings from three-leaf stage - heading stage can be hair ..... </ symptoms>
 <Pathogenic> the anamorph said gray Pyricularia </ pathogen>
 <Route of transmission of> germs conidia and mycelium ... </ transmission route>
 <Onset conditions> optimum temperature and humidity, rain, fog, ... </ incidence conditions>
<Control methods
    <The agricultural control> (1) selection of ones with huge yield and disease resistance goods ... </ agricultural control
    <Chemical control> (1) 20% tricyclazole wettable powder ... </ chemical control
</ Prevention methods
<Distribution> </ range of the distribution of the country's rice area>
<Resistant varieties validation ID = "Chongqing trial rice 2008005">
 < Cultivar name> 陵优 2 < cultivar name>
 <Varieties Source> 陵 1A× 涪 恢 9802 </ varieties SOURCE
 <Characters> The combinations are mature Indica three-line hybrid water ... </ Characteristics>
 <Rice quality performance> brown rice rate of 82.6%, and 63.6% of the milled rice rate, ... </ rice quality performance
 Yield performance of "Chongqing Rice two years trial, the average yield of 560.2kg on ... </ Yield performance>
</ Resistant varieties
</ Disease>
<Disease ID = "D0002">
<Name>
 Chinese name "rice sheath blight </ Chinese name>
      .....
</ Disease>
</ Pest weeds Data>
```

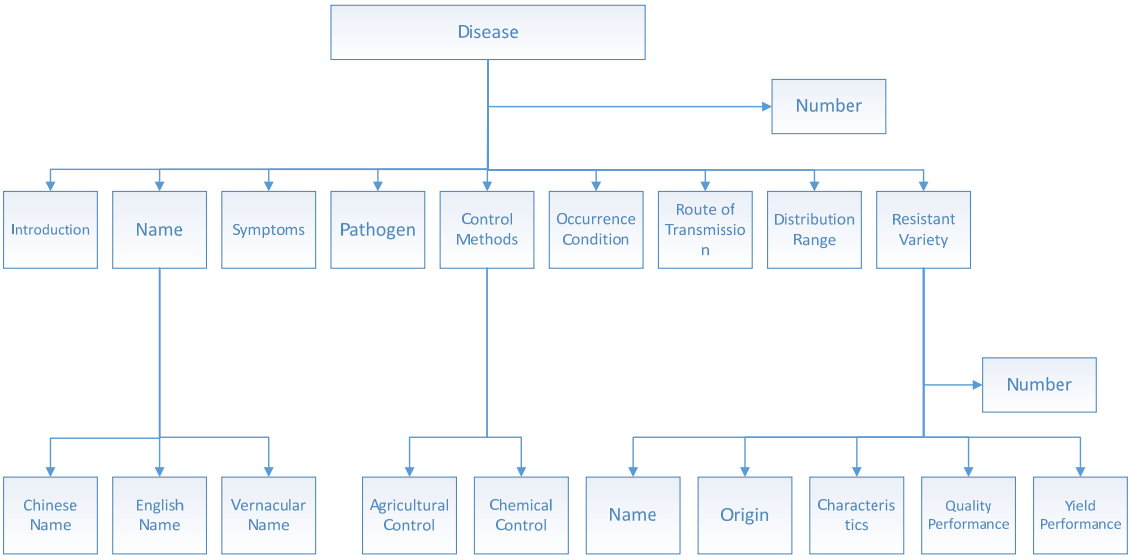

Fig. 1 (a). Rice disease tree.

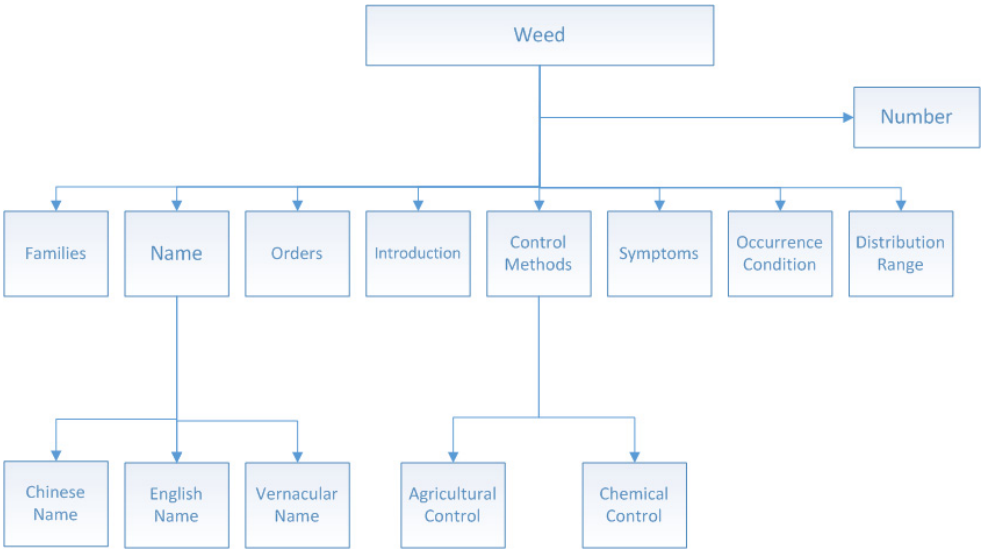

Fig. 1 (b). Rice weed tree.

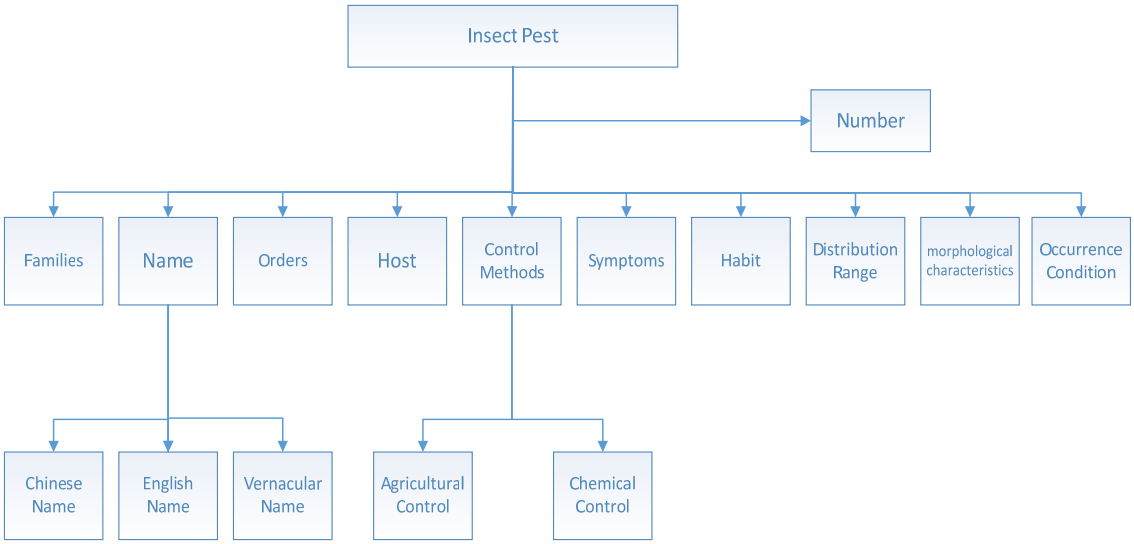

Fig. 1 (c). Rice insect pest tree.

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <xs:element name="病虫草害数据" type="DPWType"/>
    <xs:complexType name="DPWType">
        <xs:sequence>
            <xs:element name="病害" maxOccurs="unbounded">
                <xs:complexType>
                    <xs:complexContent>
                        <xs:extension base="diseaseType">
                            <xs:attribute name="编号" type="xs:string" use="required"/>
                        </xs:extension>
                    </xs:complexContent>
                </xs:complexType>
            </xs:element>
            <xs:element name="虫害" type="pestsType" maxOccurs="unbounded"/>
            <xs:element name="草害" type="weedsType" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="diseaseType">
        <xs:sequence>
            <xs:element name="名称">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element name="中文名" type="xs:string"/>
                        <xs:element name="英文名" type="xs:string"/>
                        <xs:element name="俗名" type="xs:string"/>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
            <xs:element name="简介" type="xs:string"/>
            <xs:element name="症状" type="xs:string"/>
            <xs:element name="病原" type="xs:string"/>
            <xs:element name="传播途径" type="xs:string"/>
            <xs:element name="发病条件" type="xs:string"/>
            <xs:element name="防治方法">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element name="农业防治" type="xs:string"/>
                        <xs:element name="药剂防治" type="xs:string"/>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
            <xs:element name="分布范围" type="xs:string"/>
            <xs:element name="抗病品种" maxOccurs="unbounded">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element name="品种名称" type="xs:string"/>
                        <xs:element name="品种来源" type="xs:string"/>
                        <xs:element name="特征特性" type="xs:string"/>
                        <xs:element name="米质表现" type="xs:string"/>
                        <xs:element name="产量表现" type="xs:string"/>
                    </xs:sequence>
                    <xs:attribute name="审定编号" type="xs:string"/>
                </xs:complexType>
            </xs:element>
        </xs:sequence>
    </xs:complexType>
    <xs:complexType name="pestsType">
        ……
```

Fig. 2. Rice disease XML Schema fragment.

3.Retrieval System Framework

Crop pests XML-based data retrieval system can base on user's concerning about the type of crops (rice, wheat, peanuts, soybeans, etc.) and user's identity (experts and scholars, ordinary users) to narrow the search range, effectively improve retrieval efficiency. The search results are processed in accordance to the identity of different users, returning to different results. Thus return all relevant tag information of an insect pest data to experts and scholars, and for ordinary users, only return them useful marker information. As the XML document is a text document, and inconvenient for the user to view and find, the system returns the result of the XML document and convert it to a table view to facilitate the user's browser, reach display optimization. Adding a user-defined dictionary, the key word after user authentication can be saved to the dictionary to prepare for the next search, this intelligent way will ensure the retrieval model to use the longer, get better retrieval results . The overall framework of the retrieval system is shown in Fig. 3.

Retrieval system framework process is as follows: Use DOM to analyze the data (crop diseases, insect pests and weeds) represented by the XML, and create a structure show index XML tree structure in navigation, allow users to be able to see the structure of the XML document; the cache heap area will parse the XML document data set to construct a two-dimensional table, use XML tag name as the column names in the table, allow users change the retrieval process like this: when users retrieve an XML document data, it can be converted into the retrieval of two-dimensional table.

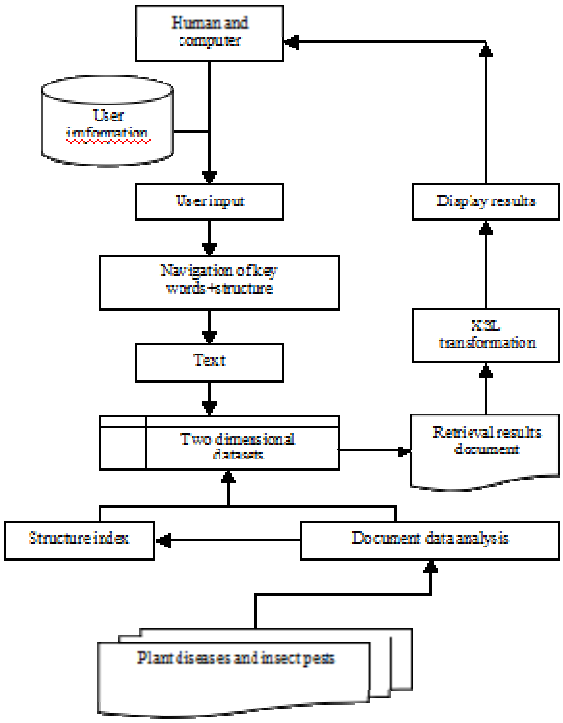

Fig. 3. Overall framework of the retrieval system.

When a user enters search keywords and select appropriate tag node in the structure of navigation, the system first judge by the user's information database, to confirm the type of crop concerned and the user's identity , narrow the range of the retrieved data sets, then conduct cutting word operation on search keywords entered by the user, get one or more criteria; Find and match for each keyword in a two-dimensional table data set, return to the corresponding two-dimensional table containing keywords information; using XSL style sheet to convert search results document described by XML into the form of a table view, showed in user-friendly browser.

4. Key Technical Analysis and Achieve

4.1. Chinese Word Segmentation

Currently, the technology of Chinese Word Segmentation is relatively mature, there are three categories: Word Segmentation based on word matching, Word Segmentation based on the statistical words, Word Segmentation based on the understanding of the cutting words. This paper use common dictionary-based matching method. In matching algorithms, Traditional dictionary mostly use txt file as storage file system, each word match need to find the entire dictionary, average seek times>n/2 (n is Dictionary length), the time complexity is O (n). In the retrieval system, every word in the dictionary is loaded into the memory of a tree, get the formation of the dictionary tree [8], so that the cut word matching operation change into a searching term in the trees, increase the speed of the cutting words.

Specific algorithm thought of Retrieval system keywords cutting:

The first step, sequential scan dictionary file, sequence of each word is configured into a tree. Take "China", "China People", "deck" as example, the tree structure is shown in Fig. 4. The figure shows that the main purpose of the contribution is that using the word in the dictionary to associate relations tree between sub trees.

The second step, when the user input keywords, word cutting program conduct matching of the first word of the keywords in the tree, if found, make it as the root node of the word, continue to find their child nodes. If the second key word is found, two words will be combined into one word, and make the second word as the root node, then sequential search for the child node, if found, these three words be combined into one word, continue until the completion of the match of the keywords in each word. If not found, then make a single word as a keyword. In order to describe more intuitive and clear, take "rice febrile disease" as example, first find the node where "rice" locate, and then use "rice" as the root node, sequential search if there is a child node of the "febrile" in many of its child nodes, if exist ,make "febrile" as root node, sequential search if there is a "disease" child node, and finally return to the word "rice febrile disease".

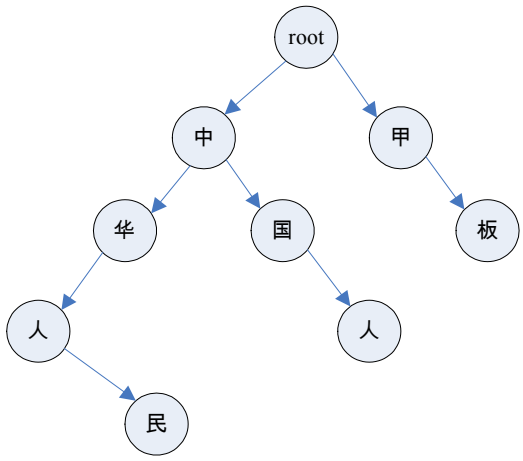

Fig. 4. Dictionary structure tree.

4.2. Implements of Retrieval

4.2.1. Two-dimensional Table Data Set

Create a structured index retrieval system by reading the XML document data, a heap area is built in the cache, and automatically build a two-dimensional table data set to store XML document, the two-dimensional table use marked tag of the source XML document as a table column name. According to the nested relationship between the XML document markup, an XML document can build interrelated multi two-dimensional table. So that retrieval of XML document data is transformed to the retrieval of the data set of the two-dimensional table. The benefits of building a two-dimensional table data set is:

1) After a user submitting a search key, the system does not have to visit source XML document data sets, and will not destroy the contents of the source XML document data, which ensures the security of the data of the XML document.

2) The user does not have to go to retrieve the text form of the XML document; the user has direct access to structured data sets of the two-dimensional table to improve the retrieval efficiency.

3) After the search is finished, the system automatically releases the heap space occupied by the data set of the two-dimensional table, avoid the waste of resources and reduce the memory space.

Fig. 5 - Fig. 8 are four tables that system builds for rice disease data. Four tables built in the cache, are virtual column table view. They are associated by "disease _Id" , all combined column names are all the elements mark rice disease represented by XML.

病害_Id	简介	症状	病原	传播路径	发病条件	分布范围	编号
0	主要为害叶...	秧苗自三叶...	无性态称灰梨孢	病菌以分生...	适温高湿，...	分布在全国...	D0001
1	此病是水稻...	一般在分蘖...	担子菌亚纲...	病菌主要以...	菌核数量是...	我国华南、...	D0002
2	水稻病害之...	水稻主要病...	稻黄单胞菌...	(1)初侵染...	带菌种子、...	除新疆外，...	D0003
3	水稻白绢病...	主要危害晚...	齐整小核菌	病菌可在土...	(1)施用的...	我国双季稻...	D0004
4	病叶由叶尖...	从秧苗期至...	稻干脐蠕孢	病菌以菌丝...	病菌以菌丝...	全国各稻区...	D0005
5	水稻矮缩病...	病叶症状表...	水稻矮缩病毒	(1)传播。...	(1)毒源。...	主要分布在...	D0006
6	一般发病田...	(1)苗期。...	串珠镰孢菌	(1)越冬。...	(1)发病期...	全国各地均...	D0007
7	水稻条纹叶...	发病之初在...	水稻条纹叶...	病毒只能通...	(1)毒源。...	我国的江苏...	D0008
8	水稻上常发...	水稻穗颈瘟...	无性态称灰梨孢	病菌以分生...	如带菌种子...	全国各地均...	D0009
9	一般发病率...	仅在穗部发...	稻绯核菌	(1)病菌以...	(1)品种因...	全国各地均...	D0010
10	叶片染病初...	叶片和叶鞘...	稻尾孢	(1)越冬。...	(1)气候因...	在全国各稻...	D0011

Fig. 5. Two-dimensional table constructed by rice disease data - Disease Table.

中文名	英文名	俗名	病害_Id
稻瘟病	Rice blast	稻热病,火烧瘟	0
水稻纹枯病	Rice sheath blight	花脚杆、烂脚杆	1
水稻白叶枯病	Rice bacterial leaf blight	看风，过火风，白叶瘟	2
水稻白绢病	Sclerotium rolfsii		3
水稻胡麻斑病	brown spot disease resistance	水稻胡麻叶枯病	4
水稻矮缩病	Rice dwarf	水稻普通矮缩病、普矮、青矮等	5
水稻恶苗病	Rice bakanae disease	水稻徒长病	6
水稻条纹叶枯病	Rice stripe disease	水稻上的癌症	7
水稻穗颈瘟	Rice Panicle Blast		8
稻曲病	Rice false smut	黑穗病、绿黑穗病、谷花病、青粉病	9
水稻窄条斑病	Rice narrow brown leaf spot	水稻褐条斑病、水稻窄斑病	10

Fig. 6. Two-dimensional table constructed by rice name data - name table.

中文名	英文名	俗名	病害_Id
稻瘟病	Rice blast	稻热病,火烧瘟	0
水稻纹枯病	Rice sheath blight	花脚杆、烂脚杆	1
水稻白叶枯病	Rice bacterial leaf blight	看风，过火风，白叶瘟	2
水稻白绢病	Sclerotium rolfsii		3
水稻胡麻斑病	brown spot disease resistance	水稻胡麻叶枯病	4
水稻矮缩病	Rice dwarf	水稻普通矮缩病、普矮、青矮等	5
水稻恶苗病	Rice bakanae disease	水稻徒长病	6
水稻条纹叶枯病	Rice stripe disease	水稻上的癌症	7
水稻穗颈瘟	Rice Panicle Blast		8
稻曲病	Rice false smut	黑穗病、绿黑穗病、谷花病、青粉病	9
水稻窄条斑病	Rice narrow brown leaf spot	水稻褐条斑病、水稻窄斑病	10

Fig. 7. Two-dimensional table constructed by rice disease prevention and treatment methods - prevention method table.

品种名称	品种来源	特征特性	米质表现	产量表现	审定编号	病害_Id
陵优2号	陵1A×浙恢9802	该组合属中...	糙米率82.6...	重庆市水稻...	渝审稻2008...	0
赣晚籼923	（莲塘早/IR36）F6与“外3...	株高，单晚...	该品种商品...	该品种经过...	赣审稻2002...	1
国稻1号	中9A×R8006	该品种属...	该组合米粒...	该组合2002...	国审稻2004032	2
连粳7号	以(镇稻88×中粳8415)F3为...	株型紧凑，...	整精米率72...	2007-2008...	苏审稻201008	5
浙盐21	早籼浙9248×舟903	株高83 cm...	糙米率83*...	平均产量8...		6
新丰2号	豫粳6号×新丰9402	属常规粳稻...	出糙率83.8...	2004年省区...	豫审稻2007...	7
松粳香2号	五优A×松98-131	香稻品种。...	出糙率79.5...	2008~2009...	黑审稻2011008	8
湘优66	湘菲A×湘恢66	该品种属籼...	整精米率66...	2006年参加...	国审稻2008...	9
佳福占	佳禾早占×佳福418	该品种属迟...	糙米率、精...	该品种2001...	闽审稻2003...	10

Fig. 8. Two-dimensional table constructed by rice disease resistant varieties - varieties of disease-resistant table.

4.2.2. Retrieval Algorithm Based on Data Sets of the Two-dimensional Table

During the search, first use word-cutting to decompose key word input by user into one or more criteria, then retrieve of each keyword .If the user has no limited to the node, then they would have access to all the data sets of the two-dimensional table; If the user selects a node, then the search range is narrowed, just need to visit the two-dimensional table where the corresponding node locate. Retrieval, combine with user's database ,determine the user's identity, concern about the types of crops, etc, further narrow the range of two-dimensional table to retrieve data sets.

XML representation contains grass data and pest data. Each disease, pests or weeds are composed by fixed node: Chinese name, English name, common name, profile, symptoms, etc. Information returned from retrieval should be complete, including these nodes, but the retrieval system divide these nodes separately in a different table when building a two-dimensional table data set , we need to diaplay the completed results. Specific algorithm to return search results are the following:

For traverse and retrieve every keyword term
Read every rows of the table where key word locate in and record Id column values
For traverse each data table set from where all the two-dimensional table data set dslist in dscache heap area
in ds find the data set tables where Tern appears
For traverse tables where each table dt
Tin ds find the datarows where Tern appears
For traverse each line dr of datarows
In Id ,judge whether its Id is in accordance with Id from the For loop, if they are same ,then display information in the dr
End For
End For
End For
End For

For example, users enter a search keyword "rice febrile disease" and select the name of the node of rice under the name of Chinese, retrieve first row of data in a two-dimensional table data set from name table (Fig. 6), but we hope to show not just the name of the table of information, but also to show the other tables related to rice febrile disease, and then combine with a complete disease information. Taking into account these tables are associated with the column "disease _Id", if the row of data of the name table is known, all data sets of the two-dimensional table must be traversed, to find out all of the data table from the data set, and then from the data find all the rows in the table, and then loop through all the rows, see whether Id column we find out is in accordance with the Id column information from name table, if they are same, then display information.

4.3. Optimized Search Results

Extensible Stylesheet Language (XSL) is an XML-based language, which is designed to transform the XML document into another version of XML

document or HTML document [9]. The system returns to data fragment of source XML document, which is inconvenient for users to browse. Considering the XML document data is separated with the display format ,for the same results from XML documents, we can show a different format according to the different users. For experts and scholars, all XML markup data of the pests would be shown, for ordinary users, they are more concerned with the control methods against crop grass pest , we only display Chinese name and prevention methods. Retrieval system displays search results from XML document into a table view.

```
<?xml version="1.0" encoding="utf-8" ?>
<病害数据>
    <病害 编号="D0001">
        <名称>
                <中文名>稻瘟病</中文名>
                <英文名>Rice blast</英文名>
                <俗名>稻热病,火烧瘟</俗名>
        </名称>
        <简介>主要为害叶片、茎秆、穗部……</简介>
        <症状>秧苗自三叶期—穗期均可发……</症状>
        <病原>无性态称灰梨孢</病原>
        <传播途径>病菌以分生孢子和菌丝……</传播途径>
        <发病条件>适温高湿，有雨、雾、……</发病条件>
        <防治方法>
                <农业防治>(1)选用高产抗病品……</农业
                        防治>
                <药剂防治>(1)20%三环唑可湿性粉……</
                        药剂防治>
        </防治方法>
        <分布范围>分布在全国各稻区</分布范围>
        <抗病品种 审定编号="渝审稻2008005">
            <品种名称>陵优2号</品种名称>
                <品种来源>陵1A×涪恢9802</品种来源>
                <特征特性>该组合属中籼迟熟三系杂交水
                        ……</特征特性>
                <米质表现>糙米率82.6%，整精米率63.6%
                        ，……</米质表现>
                <产量表现>重庆市水稻两年区试，平均亩
                        产560.2kg对……</产量表
                        现>
        </抗病品种>
                ……
                </病害数据>
```

Fig. 9. Convert document XSL.

For Experts and scholars, XML documents is converted into table view, is shown in Fig. 10.

基本信息	中文名	稻瘟病
	英文名	Rice blast
	俗名	稻热病,火烧瘟
	简介	主要为害叶片、茎秆、穗部。因为害时期、部位不同分为苗瘟、叶瘟、节瘟、穗颈瘟、谷粒瘟。
	症状	秧苗自三叶期--穗期均可发生，病斑的类型、形状、色泽等因天气、品种和病斑年龄而异。叶瘟一般于分蘖期以后发生，由于品种和气象条件不同，叶瘟有四种症状， ①慢性型； ②急性； ③白点型； ④褐点型。除褐点型（抗病品种上产生）、白点型（感病品种嫩叶上产生）外，在田间常见的有急性型与慢性型病斑。
	病原	无性态称灰梨孢
	传播途径	病菌以分生孢子和菌丝体在稻草和稻谷上越冬。翌年产生分生孢子借风雨传播到稻株上，萌发侵入寄主向邻近细胞扩展发病，形成中心病株。病部形成的分生孢子，借风雨传播进行再侵染。
	发病条件	适温高湿，有雨、雾、露存在条件下有利于发病。菌丝生长温限8-37℃，最适温度26-28℃。孢子形成温限10-35℃，以25-28℃最适，相对湿度90%以上。
	分布范围	分布在全国各稻区
防治方法	农业防治	(1)选用高产抗病品种(2)加强栽培管理(3)及时处理病稻草并进行种子消毒
	药剂防治	(1)20%三环唑可湿性粉剂75～100克。(2)40%稻瘟灵(富士一号)乳油75～100毫升。(3)2%加收米(春雷霉素)液剂80～100毫升。(4)21．2%加收热必可湿性粉剂75～100克。(5)13%稻洁(春•三环)可湿性粉剂130～150克。兑水50公斤喷雾
抗病品种	审定编号	渝审稻2008005
	品种名称	陵优2号
	品种来源	陵1A×涪恢9802
	特征特性	该组合属中籼迟熟三系杂交水稻。作中稻种植全生育期154d比对照Ⅱ优838短1d株高102.5cm栽秧1.2万窝的 陵优2号
	米质表现	糙米率82.6%，整精米率63.6%，粒长6.7mm，长宽比3.0，垩白粒率6%，垩白度1.2%，胶稠度86mm，直链淀粉含量22.0%，达到国家《优质稻谷》标准(2)级

Fig. 10. XML table view.

5 Conclusions

XML has become main standard for expressing and exchanging data on the Internet, how to conduct fast and efficient retrieval of XML data, it is one of the main problems we face. This paper, combine with the characteristics of the XML data, use the method that system automatically builds a data set of the two-dimensional table, and take crop pests data for example, build an crop pests data retrieval system XML-based. The retrieval system avoids the overhead of time and space which are required to create a content index .For element tags which can identify an insect pest data , such as the Chinese name, English name, common name, etc, make a strict interpretation of their structural constraints during retrieval, the accuracy rate of results from retrieval can reach 100 %. Retrieval system uses the most simple string matching method, which is slow. Although in the current data sets case, the disadvantage of this algorithm hasn't been reflected, however, with the incensement of the data set, the algorithm remains to be further optimized. In addition, the retrieval system does not consider the order of presentation of search results, which is the next focus of the work.

Acknowledgements

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues. This work was supported by the National Natural Science Foundation of china (Grant No. 31371533) . This work was also supported by the Natural Science

Foundation of Anhui Province (Grant No.1308085MF89) and by Key Technologies R & D Program of An Hui Province (Grant No.1308085MF89).

References

[1]. Abiteboul S., On views and XML, *ACM SIGMOD Record*, 28, 4, 1999, pp. 1-9.

[2]. Burg M, Wong R. K., Structural proximity searching for large collections of semi-structured data, in *Proceedings of the Int. Con. On Information and Knowledge Management (ACMCIKM),* 2001, pp. 175-182.

[3]. Seo, D. M., Yoo, J. S., Cho, K. H., An Efficient XML Index Structure with Bottom-Up Query Processing, in *Proceedings of the ICCS*, 2007, pp. 813-820.

[4]. Aggarwal C., Yu P. S., A survey of uncertain data algorithms and applications, *Knowledge and Data Engineering,* 21, 5, 2009, pp. 609-623.

[5]. Zhang Z. P., Li R., Cao S. L., Zhu Y. Y., Similarity metric for XML documents, in *Proceedings of the Workshop on Knowledge and Experience Management (FGWM'03),* Karlsruhe, 2003, pp. 255−261.

[6]. Flesca S., Manco G., Masciari E., Pontieri L., Pugliese A., Detecting structural similarities between XML documents, in *Proceedings of the Int. Workshop on the Web and Databases (WebDB),* 2002, pp. 55−60.

[7]. Chehreghani. M. H., Rahgozar, M., Lucas, C., Mining Maximal Embedded Unordered Tree Patterns, in *Proceedings of the IEEE Symp. on Computational Intelligence and Data Mining*, 2007, pp. 437-443.

[8]. Kimelfeld B, Sagiv Y., Matching twigs in probabilistic XML, in *Proceedings of the 33rd Int Conf on Very Large Data Bases,* 2007, pp. 27-38.