INFSCI 2927

Independent Study in Application of Information Technology

INDEPENDENT STUDY REPORT

ON

DOCUMENT SIMILARITY DISTANCE VISUALIZED TOOL

Term: Fall 2005

Mr.Kitipong Techapanichgul

INFSCI 2927

Independent Study in Application of Information Technology

PROJECT OVERVIEW

- **Course: INFSCI 2927 Independent Study in Application of Information Technology**
- **Term: Fall 2005**
- **Advisor: Dr. Peter Brusilovsky**
- **Topic:** Document Similarity Distance Visualized Tool

OBJECTIVE

To develop a visualized tool for simulating a document space, which represents the similarity among documents. The tool can present the space in both 2D and 3D graphics, so common users are able to perceive how the document space looks like easily. For expert users, it can be used to display the distance data according to the give data source.

BRIEF DESCRIPTION

This tool was initiated and developed as a part of the Information Storage and Retrieval course mainly concerning about the processes of documents' similarity distance calculation. The primary goal of the tool was to provide users understanding the calculation processes. Fortunately, since the tool seems to be useful for various projects involving with information retrieval area, it has been decided to intensively rebuild the visualization part in order to serve other existing systems as a visualized tool. So, it can be add, modified, upgrade to be support the other systems better. This current version can be used as both standalone and visualized service depending on the availability of the data source. Using it as a visualized service will be discussed more in the following topics.

TERM EXPLANATION

Document space is the representation of the set of documents. There can be different ways of representing documents, however mostly it is used to show the distance or the similarity between different documents. When documents are close each other in the document space this represents that they are similar or relevant document in the set. When documents are far apart each other in the document space this means they are not similar or relevant each other. Therefore, to represent documents in document space, the distances between documents are necessary.

Document distances can be calculated with several distance measurement methods. Mainly used methods are vector-based such as Euclidean measure and Cosine measure.

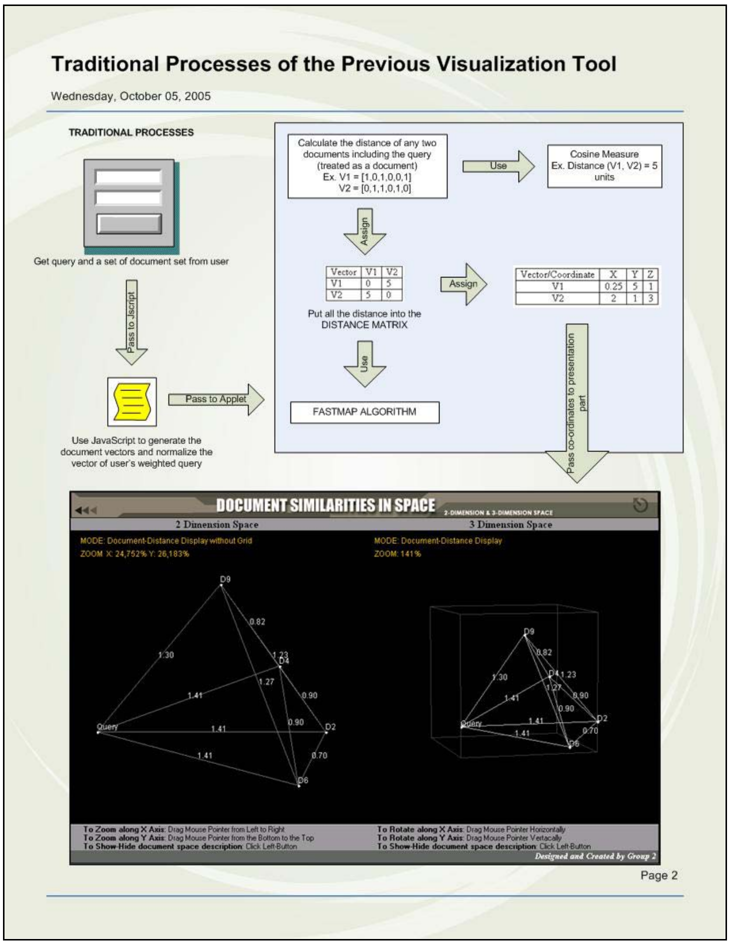
In vector-based models, documents and query are represented with their weights which correspond to the importance of the term in the document. Here in our project, term weights are related to the frequency of the terms in the document. Each term in the document will have a vector value, by which the distace between documents will be calculated. *The vector* is calculated as following formula for the first term in a document, for example.

Tool OVERVIEW

Pre-Research on Information Visualization Application

According to the paper written by Gary Geislar (this paper can be found at http://www.ils.unc.edu/~geisg/info/infovis/paper.html), it support the idea on making the document space as two dimension space in order to increase users' perception on the distance. Two dimensions representation has been widely used a wonderful invention like Maps to clearly present the distance among points in space. Also, three dimensions space is another way to present the objects in real world. Because all humans are familiar in the 3 dimensions space, the 3 dimensions document space brings users more understanding and powerful presentation about the spatial relationship.

Traditional Processes of the Previous Visualization Tool



Traditional Processes of the Previous Visualization Tool

Wednesday, October 05, 2005

TRADITIONAL PROCESSES

Get query and a set of document set from user

Use JavaScript to generate the document vectors and normalize the vector of user's weighted query

Calculate the distance of any two documents including the query (treated as a document)
Ex. V1 = [1,0,1,0,0,1]
V2 = [0,1,1,0,1,0]

Cosine Measure
Ex. Distance (V1, V2) = 5 units

Vector	V1	V2
V1	0	5
V2	5	0

Put all the distance into the DISTANCE MATRIX

Vector/Coordinate	X	Y	Z
V1	0.25	5	1
V2	2	1	3

FASTMAP ALGORITHM

MODE: Document-Distance Display without Grid
ZOOM X: 24,752% Y: 26,183%

MODE: Document-Distance Display
ZOOM: 141%

To Zoom along X Axis: Drag Mouse Pointer from Left to Right
To Zoom along Y Axis: Drag Mouse Pointer from the Bottom to the Top
To Show-Hide document space description: Click Left-Button

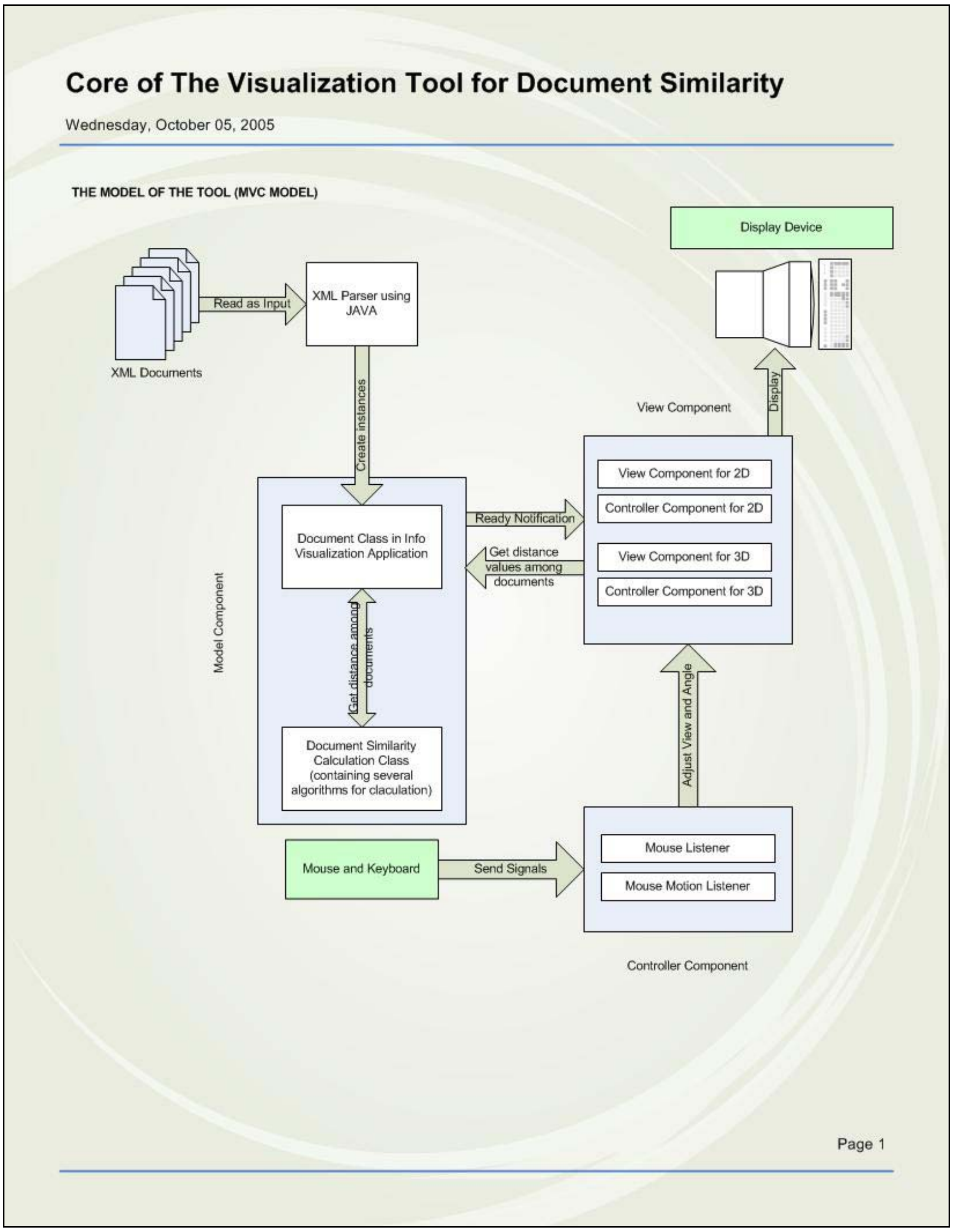
To Rotate along X Axis: Drag Mouse Pointer Horizontally
To Rotate along Y Axis: Drag Mouse Pointer Vertacally
To Show-Hide document space description: Click Left-Button

Designed and Created by Group 2

Page 2

From the chart beside, it shows the primary processes of the previous tool. It receives query and a set of document from users. Then, it use JavaScript to calculate the query vector as well as document vectors. After that it sends those values into the java part, which will use cosine measure technique to calculate the distance among the documents. Finally, FastMap algorithm will be used to calculate the position of each document in space and then present them in both 2D and 3D space. The detailed documentation about the previous system can be found at http://www.sis.pitt.edu/~ktech/IR_Project/index.html.

Tool Design



Core of The Visualization Tool for Document Similarity

Wednesday, October 05, 2005

THE MODEL OF THE TOOL (MVC MODEL)

XML Documents

Read as Input

XML Parser using JAVA

Create instances

Model Component

Document Class in Info Visualization Application

Get distance among documents

Document Similarity Calculation Class (containing several algorithms for claculation)

Ready Notification

Get distance values among documents

View Component

View Component for 2D

Controller Component for 2D

View Component for 3D

Controller Component for 3D

Display

Display Device

Adjust View and Angle

Mouse and Keyboard

Send Signals

Mouse Listener

Mouse Motion Listener

Controller Component

Page 1

From the chart above, the new tool was designed based on MVC (Model-View-Controller) model and has been developing by Java technology because this model and technology are capable of tool extensibility as well as flexibility. The model component primarily works as space management who holds all the data required by space displayer to present them in nice graphic format. At this point, the main ".java" files containing in the model component will be explained as follow:

FileName	Main Responsibilty
Document	Define structure of a document
DocumentInSpace	Define Structure of a document when is in document space (inherited from Tool's Document Object)
FastmapModelManager	Implement the methods for SpaceCoordinatorManager
InputProvider	Get the XML document from the specified path (URL)
SpaceModel	Hold the entire data used for calculating the new position according to the user input. Also, it supports necessary data to other components.
XMLParser	Parse the XML document received from the specified path

From the model, it needs some kinds of viewers to implement and work as the space displayer. Besides, since the tool is intended to serve various groups of users. The developer makes the tool available online by implementing it as Applet which the controller component was integrated inside. So, the main ".java" files containing in the view component and controller component will be explained as follow:

FileName	Main Responsibilty
AppletMenuBar	Manage about the menu presentation in standalone mode
AppletSoundList	Collect sounds used in the system to improve tool interactivity
AppletSoundLoader	Load the sound from the package
Transparency	Make the image of file icon transparent in the space
MyFileTransferHandler	Support the tool usability by allowing user to drag and drop the legal file (txt file and .xml file)
SpaceDisplay	Hold all the view components together and manage all the view processes

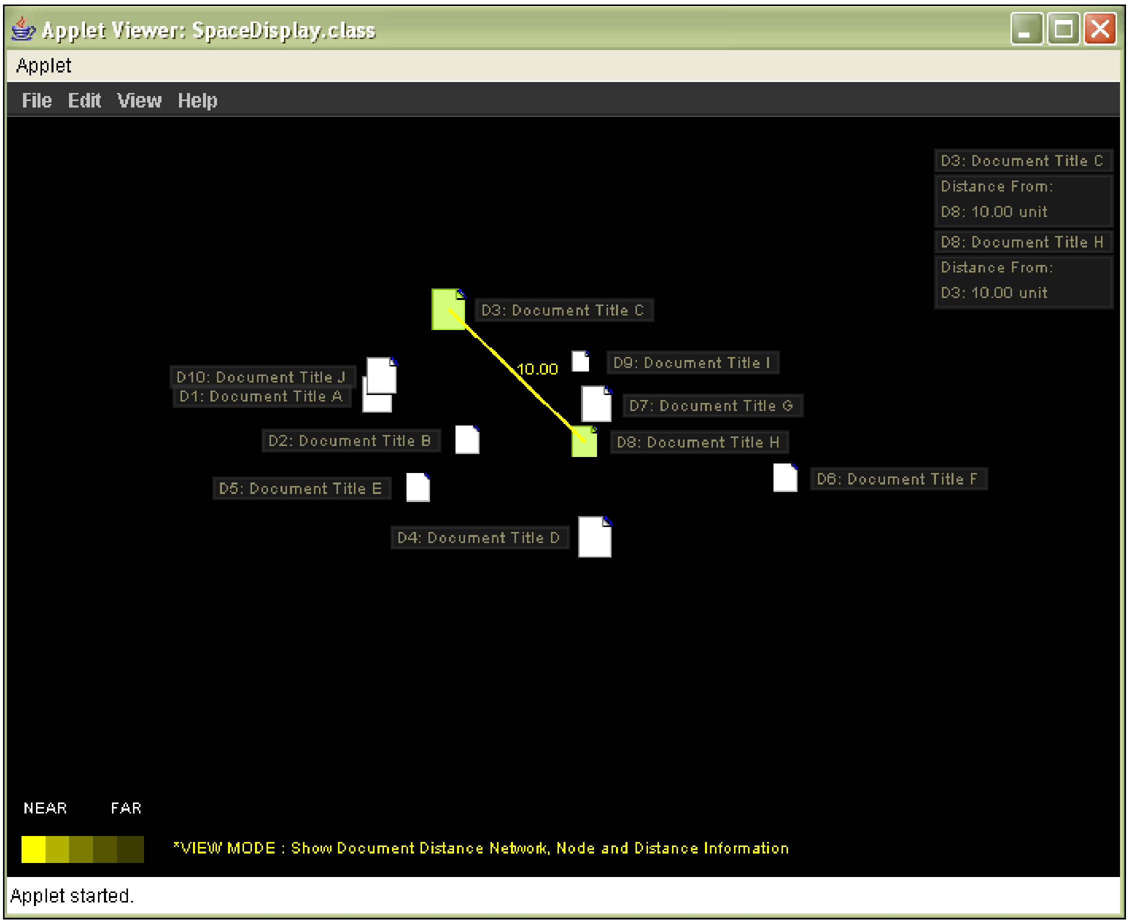
Tool Extended Features

Although it seems that there is no big change in the tool's interface at the first glance, the fact is that the internal system has been rebuilt entirely to improve the tool extensibility. Also, the presentation is so delicate in this version. To begin with the document presentation, the document has been displayed as icon indicating file types. Furthermore, the size of the objects depends on how they far from the users' eyes. Furthermore, the document that is located in front of the other documents will be painted at last to improve the perception ability and delicate presentation. The color opacity of lines representing the distance among the documents will be more intense. Specifically speaking, it depends on how close between the two documents. Besides, the tool's usability has been enhanced by allowing user to drag and drop the distance file (both text and xml file) to the applet. Also, users can only click the focused documents to see the distance and information about the document without flooding with unwanted information. To help users see the overall of the improvement, the extended features is summarized in the table below.

Extension Tool Feature	Feature Description
Interactivity	
• Adjustable Space View	User can see around the space by simply moving mouse to control the space movement
• Interested Document Highlight	User can click on the interested document to see the document information and the distance among selected documents
• Floating Document information	User can see the brief information about the document by simply move the mouse over the interested document
Usability	
• Drag and Drop Support	User can drag the distance data file to the space without using the tool menu bar
Readability	
• Network Density	The density of lines in the tool depends on how documents close to each other. It can help users to figure it out faster when it has a dense network
• File Type Recognition	The icon used to present a file type can increase information readability in overall
• Space Depth improvement	The space calculates Z space according to mouse movement which doesn't appear in the previous version
Integrated ability	
• Both mode in the same tool	The tool can be used as a service by specifying the URL of the data source (only xml file is accepted). Also, if it can't find a proper data source, it will change its mode to be stand alone mode, which provides users more tool features
Additional Features	
• Online Help	The manual of the toll is provided online, so users can check it out easily

CURRENT INTERFACE

To bring readers an idea what the tool look likes, the screenshot of the current interface is provided below. Also, interested researches can use with this applet via the following URL: http://www.sis.pitt.edu/~ktech/DocumentSpaceProject/DocumentSpace.htm



D3: Document Title C
Distance From:
D8: 10.00 unit

D8: Document Title H
Distance From:
D3: 10.00 unit

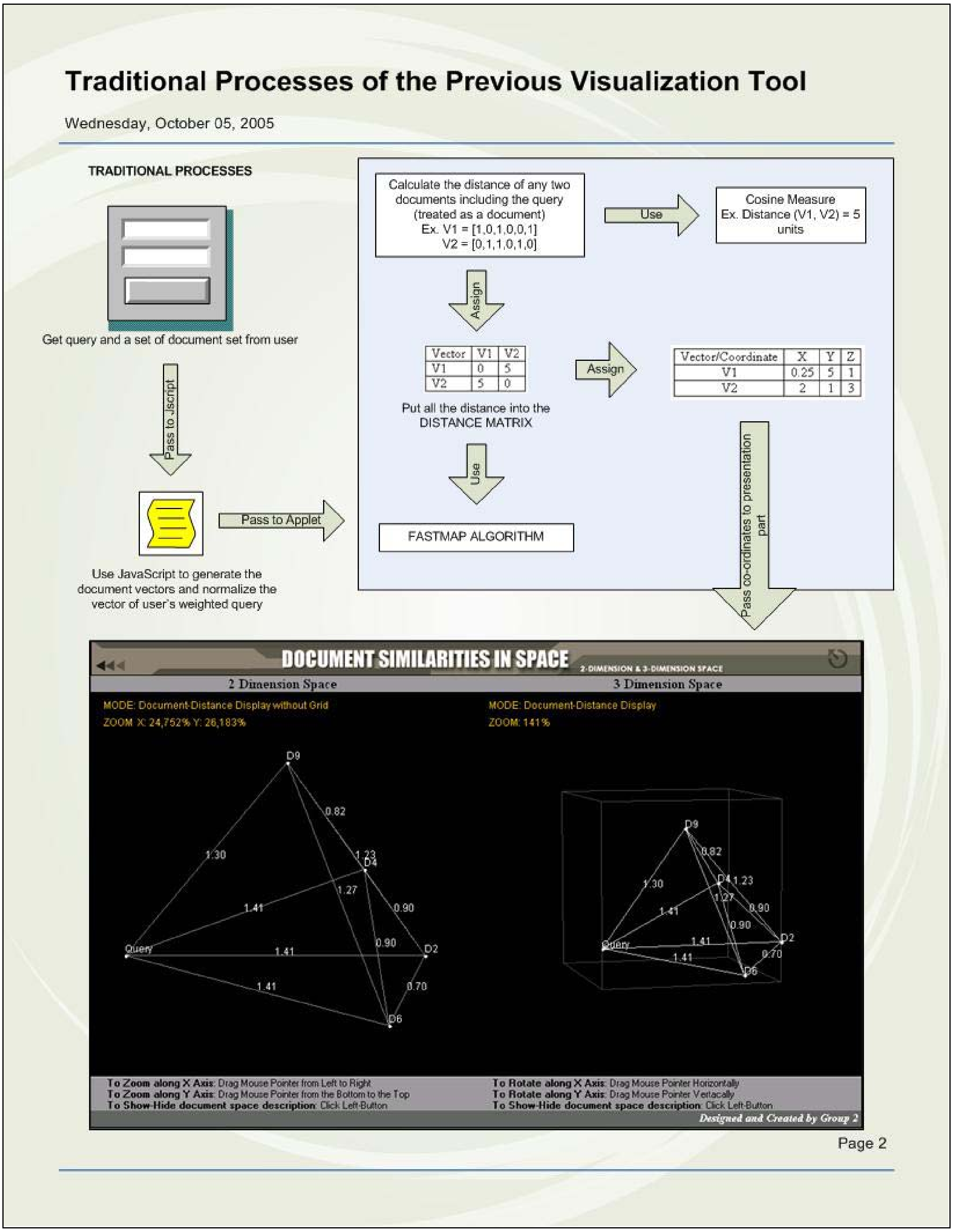
*VIEW MODE : Show Document Distance Network, Node and Distance Information

Applet started.

FUTURE IMPROVEMENTS

- Appropriate algorithm Integration for similarity distance calculation
- Add on VRML exporting feature
- Document Searching Functions (by using document information or document content)
- Reducing the processing cost for a large collection of documents

Appendix A: Figures in Full



Traditional Processes of the Previous Visualization Tool

Wednesday, October 05, 2005

TRADITIONAL PROCESSES

Get query and a set of document set from user

Pass to Jscript

Use JavaScript to generate the document vectors and normalize the vector of user's weighted query

Pass to Applet

Calculate the distance of any two documents including the query (treated as a document)
Ex. V1 = [1,0,1,0,0,1]
V2 = [0,1,1,0,1,0]

Use

Cosine Measure
Ex. Distance (V1, V2) = 5 units

Assign

Vector	V1	V2
V1	0	5
V2	5	0

Put all the distance into the DISTANCE MATRIX

Assign

Vector/Coordinate	X	Y	Z
V1	0.25	5	1
V2	2	1	3

Use

FASTMAP ALGORITHM

Pass co-ordinates to presentation part

DOCUMENT SIMILARITIES IN SPACE

2-DIMENSION & 3-DIMENSION SPACE

2 Dimension Space

MODE: Document-Distance Display without Grid

ZOOM X: 24,752% Y: 26,183%

3 Dimension Space

MODE: Document-Distance Display

ZOOM: 141%

To Zoom along X Axis: Drag Mouse Pointer from Left to Right
To Zoom along Y Axis: Drag Mouse Pointer from the Bottom to the Top
To Show-Hide document space description: Click Left-Button

To Rotate along X Axis: Drag Mouse Pointer Horizontally
To Rotate along Y Axis: Drag Mouse Pointer Vertacally
To Show-Hide document space description: Click Left-Button

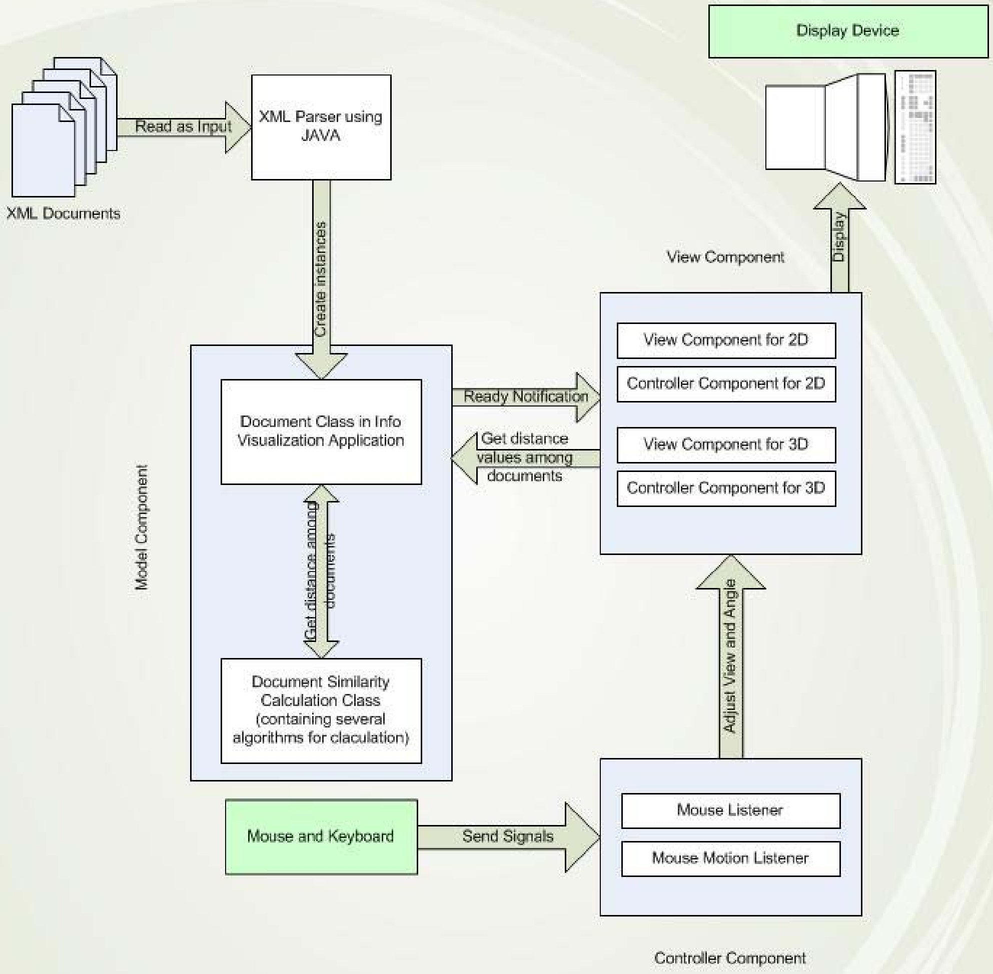
Designed and Created by Group 2

Page 2

Core of The Visualization Tool for Document Similarity

Wednesday, October 05, 2005

THE MODEL OF THE TOOL (MVC MODEL)



Display Device

XML Documents

Read as Input

XML Parser using JAVA

Create instances

Model Component

Document Class in Info Visualization Application

Get distance among documents

Document Similarity Calculation Class (containing several algorithms for claculation)

Ready Notification

Get distance values among documents

View Component

View Component for 2D

Controller Component for 2D

View Component for 3D

Controller Component for 3D

Display

Adjust View and Angle

Mouse and Keyboard

Send Signals

Mouse Listener

Mouse Motion Listener

Controller Component

Page 1