

Virtual Fashion -

Tracking and Analyzing Cultural Dispersion on the World Wide Web

Ta-gang Chiou

B.S. Electrical Engineering
National Taiwan University
June 1998

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
In partial fulfillment of the requirements for the degree of
Master of Science in Media Technology
At the
Massachusetts Institute of Technology

June, 2000

©Massachusetts Institute of Technology, 2000.
All rights reserved.

author

Ta-gang Chiou

Program in Media Arts and Sciences
May 1, 2000

certified by

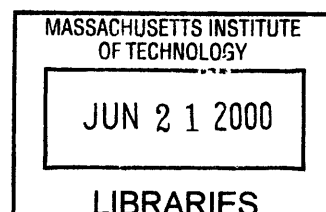
Judith S. Donath

Assistant Professor of Media Arts and Sciences
Thesis Supervisor

accepted by

Stephen A. Benton

Chair, Departmental Committee on Graduate Studies
Program in Media Arts and Sciences



ROTCH

Virtual Fashion -

Tracking and Analyzing Cultural Dispersion on the World Wide Web

Ta-gang Chiou

B.S. Electrical Engineering

National Taiwan University

June 1998

Submitted to the

Program in Media Arts and Sciences,

School of Architecture and Planning,

In partial fulfillment of the requirements for the degree of

Master of Science in Media Technology

At the

Massachusetts Institute of Technology

June, 2000

Abstract

In the real world, people clothe themselves in garments whose cut and design encodes information about their social identity. This encoding changes temporally as the design spreads throughout a population: this is the basis of "fashion." A similar sense of fashion has emerged on the World Wide Web (WWW), as people embellish their homesites with links, pictures, and other objects that exhibit similar patterns of dispersion.

I have developed tools and algorithms for tracking and analyzing this "virtual fashion." The initial approach is to examine a set of selected homesites each week and track the spread of links. By developing a system for collecting and analyzing the data, this research provides both macro and micro readings of the phenomenon of virtual fashion. The system shows what is popular, ways that things are related, and what is emerging online. I also use data collected by the system to think about existing social theories of fashion and see how they may help develop models of virtual fashion. This research helps people further understand how the WWW functions as a social environment.

Thesis Supervisor: Judith S. Donath

Title: Assistant Professor of Media Arts and Sciences

Virtual Fashion -

Tracking and Analyzing Cultural Dispersion on the World Wide Web

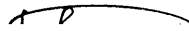
Ta-gang Chiou

The following people served
as readers for this thesis:

reader

Pattie Maes

Associate Professor, Software Agents Group
MIT Media Laboratory

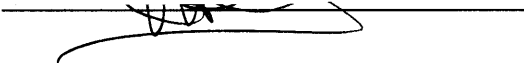


reader

Henry Jenkins

Associate Professor of Literature
MIT

MIT



Acknowledgements

I am grateful for the help of many people who have made this work possible.

I really appreciate the support I have received from my advisor, Professor Judith Donath. Judith has always provided great input to this research. Her keen insight into online social spaces has always been of great help to me.

I would like to thank the readers of this thesis, Professor Pattie Maes and Professor Henry Jenkins, for their ideas and encouragement. I would also like to thank Wen-Jiuan Tsai, who has been constantly advising me and helping me in various ways.

I thank two good undergraduate researchers, Kan Liu and Constantin Chiscanu. Kan did a great amount of work in developing the visualization program. We have tried several failed visualizations and have finally come up with a working one. Without Kan's help, it would have taken me even more time for me to finish the system. Constantin contributed largely to the section on the low-level model of cultural transmission with his expertise in mathematics.

I am thankful to my colleagues at the Sociable Media Group for their unfailing willingness to discuss this project and for their insightful comments: Fernanda Viegas, Roy Rodenstein, Karrie Karahalios, Joey Rozier, Rebecca Xiong, Dana Spiegel, Matthew Lee, and Danah Beard. Each of them has different expertise and I have learned a great deal from them.

My deep thanks to Daniel Diemers' comment on the research. He carefully read through my paper and gave me lots of insight into the research. If I am going to continue this work

as a serious sociological study, the references he gave me will be important groundwork.

I am appreciative of Ching-Huang Yuan's valuable advice on data mining.

I am thankful to Andreas Weigend. He gave me hints on mathematics that is helpful to the project.

I also appreciate Benjamin Vigoda, Michael Best, Ken Kung, Erik Blankinship, Davie Hong, and Susan Spilecki for their input to this research.

The MIT Media Lab is a fabulous place where open-minded experts in different fields work together and inspire one another. I am grateful to Professor Nicholas Negroponte for creating this lab.

The help from my parents is beyond description. Without their unconditional support, this thesis could never have been done.

Extended Abstract

In the real world, people clothe themselves in garments whose cut and design encodes information about their social identity. This encoding changes temporally as the design spreads throughout a population: this is the basis of "fashion." A similar sense of fashion has emerged on the World Wide Web (WWW), as people embellish their homesites with links, pictures, and other objects that exhibit similar patterns of dispersion. By following the rise and fall of these trends, we may learn a great deal about the structure of a society, for these patterns of cultural dispersion delineate subcultures, provide evidence of individual's role and status, and mark the flow of information.

My goal is to follow the diffusion of virtual fashion over time and analyze it. The initial approach is to examine a set of selected homesites each week and track the spread of virtual objects in the form of links. I have developed a system and algorithms for tracking and analyzing this "virtual fashion." The system can reliably track the spread of objects, efficiently eliminate the noise, smoothly integrate the change to show the long-term trend, and effectively delineate sub-cultural hierarchies within the virtual community based on the diffusion of virtual objects. It shows what is popular, ways that things are related, and what is emerging online. A sample spatial-temporal archive of about eight thousand homesites has been collected during a one-year period. Many interesting case studies have been done based on the archive.

The idea of tracking the popularity change and diffusion pattern of objects among homepages over time is a new one. Although there are several existing Web-based data-mining services based on hit-rate, link information, or textual context, none of them shows the temporal dynamics of the popularity of virtual objects, nor do they reveal the role individuals play in the diffusion of online culture. In contrast, my system may help

find out which homepage adopts which popular object at what time and for how long. This information lays the groundwork of further research in theoretical modeling of people's interaction pattern on homepages.

I have used data collected by the system to think about existing social theories of fashion and see how they may help develop models of virtual fashion. For example, analysis on the data set shows that the hierarchy of homesites regarding the diffusion of virtual fashion is as follows:

1. There are different groups of homesites. They form different sub-cultures of interest. Members of each group may also change their interests over time and thus leave the group.
2. The uniform and multi-layered class hierarchy described by trickle-down theory is not the case online. The "trickle-down" effect is generally weak and the social structure is fragile between homesites.
3. There are some individual-to-individual chains, but I have not found this to be a significant phenomenon.

In brief, data collected by the system brings up many deeper sociological questions about the mechanism of virtual communities.

As online community formation and other interpersonal interactions become increasingly widespread, the Web's role as a place where people establish their identity becomes more and more important. I am interested in understanding how cultures evolve in this milieu. Fashion is a key part of this evolution. This research provides both macro and micro readings of the phenomenon of virtual fashion.

Table of Contents

- 1. INTRODUCTION9**
 - 1.1 Definition of Fashion [9]**
 - 1.2 What Constitutes Virtual Fashion on the WWW [9]**
 - 1.3 Why is Examining the Spread of Virtual Fashion a Significant Problem... [10]**
 - 1.4 The Organization of This Thesis [11]**
- 2. SOCIOLOGICAL ISSUES12**
 - 2.1 Homesite as One's Online Identity [12]**
 - 2.2 Person-object Relation [12]**
 - 2.3 Fashion Cycle - From the Physical to the Virtual World [13]**
 - 2.4 Theories of Fashion and Their Implication [15]**
- 3. SYSTEM DESIGN.....20**
 - 3.1 System Overview [20]**
 - 3.2 Archiving Subsystem [21]**
 - 3.3 Extracting Subsystem [25]**
 - 3.4 Integrating Subsystem..... [26]**
 - 3.5 Clustering Subsystem [28]**
- 4. OUTPUT AND CASE STUDIES.....42**
 - 4.1 Online Service [42]**
 - 4.2 General Findings..... [51]**
 - 4.3 Case Studies [55]**
 - 4.4 Visualization..... [64]**
- 5. DISCUSSION ON THEORETICAL MODELING75**
 - 5.1 Background Research [75]**
 - 5.2 Modeling Framework..... [76]**
 - 5.3 Verification of the Trickle-down Theory [76]**
- 6. CONCLUSION81**
- 7. BIBLIOGRAPHY82**

1. Introduction

1.1 Definition of Fashion

In the real world, fashion communicates social signals from people (Polegato and Wall 1980). People clothe themselves in garments whose cut and design encodes information about their social identity. More broadly, a fashion may be a real or virtual object that spreads throughout a population and whose social meaning changes over time. The object may be meaningful by itself or not, and the meaning can be context dependent. For example, although made of the same material, the black gauze of the funeral veil means something very different from that sewn into the bodice of a nightgown (Davis 1992). Finally, the social context in which fashion diffusion occurs may determine its direction, tempo, and dynamics (McCracken 1988). For instance, casual clothing can be easily seen from one classroom to another at MIT, but not in the military. Fashion need not be clothing: an example is the popular Dilbert cartoons that get copied and dispersed through an office. The ever-changing trend of pop music is another example of fashion.

1.2 What Constitutes Virtual Fashion on the WWW

Just as people adorn themselves with clothing, they now display identity online through the presentation of homesites (Donath 1995a). On the World Wide Web (WWW), people embellish their homesites with links, pictures, sounds, etc. These items, as well as the content and overall design of the pages, are cultural features whose spread from site to site constitutes fashion on the Web. For example, if somebody puts an image on his homesite, and other people link or copy it to their own homesites, the image becomes popular among the set of homesites. The more people do this, the more popular it becomes on the WWW. When people unlink it, the popularity declines (Pirulli and

Pitkow 1997). By following the rise and fall of these trends, we may learn a great deal about the structure of a society, for these patterns of cultural dispersion delineate sub-cultures (Blumer 1969, Chiou and Donath 2000), provide evidence of individual's role and status (Simmel 1904), and mark the flow of information.

In this thesis, I define the term "homesite" to be a set of web pages, and sometimes a Web site, a certain individual owns. Homesites are usually designed to be an online self-presentation of its owner.

1.3 Why is Examining the Spread of Virtual Fashion a Significant Problem?

Fashion trends are difficult to detect both in the real world (Davis 1992) and online. Fashions follow complex trajectories making it difficult to obtain precise information. In the real world, reliable and timely information about the changes in fashion (whether in clothing, slang, music, etc.) is difficult to track. Online, although there have certainly been a number of virtual fashions even in the short history of the WWW, it has been difficult to perceive their structure and extent especially because it is difficult to obtain an overview of large groups and their changing behavior. Efficient heuristics for tracking, extracting, storing, and analyzing virtual fashion are clearly needed, especially for market analysts to understand people's change of taste, and for researchers to understand how the web functions as a social space.

As online community formation and other interpersonal interactions become increasingly widespread, the Web's role as a place where people establish their identity becomes increasingly important (Donath 1995a). I am interested in understanding how cultures evolve in this milieu. Fashion is a key part of this evolution. For example, what are the different roles of homesites? What are the different roles of fashions? What are their diffusion patterns and life cycles? With this research, I hope to further understand how the WWW functions as a social environment, as compared to the real world. This

knowledge may help people design better online tools, and indicate further areas of research.

To a certain degree, the virtual world presents unprecedented opportunities to observe and model social phenomena such as fashion in an easily quantifiable environment (Gibson et al. 1998). It is possible to know which person adopts which virtual fashion at what time and for how long. I also use data collected by the system to think about existing social theories of fashion and to see how they stand up in the virtual world.

1.4 The Organization of This Thesis

In the thesis, I first review related sociological issues. To support the rationality of the virtual fashion analogy, I give examples of how people deal with objects in the real world, and compare this behavior to people's manipulation of online objects. To understand how fashion might work, I then survey observation on change of clothing fashion, and review the two most distinguished fashion theories.

After established the analogy, I describe my system for tracking and analyzing virtual fashion, and discuss issues, challenges, and my experience in designing and implementing the system.

Then, a simple online service based on the data I collected is presented. With the data, theoretical modeling of fashion online is possible. A short introduction to my work on modeling fashion online is also given at the end of the thesis.

2. Sociological Issues

In this chapter, I briefly review sociological issues about person-object relation and fashion in the real world. These ideas help us understand how virtual fashion may work, and thus affect the direction of the project.

2.1 Homesite as One's Online Identity

The concept of social identity points to the configuration of attributes and the attitude that persons seek to and actually do communicate about themselves. This may include not only symbols of social class, but also any aspect of self about which individuals can through symbolic means communicate with others (Davis 1992).

In similar ways, a homesite is fast becoming one's online self-presentation (Donath 1995a). Readers may not only read facts about the person, but also perceive more subtle aspects of him or her.

2.2 Person-object Relation

A comparison of how people deal with real-world objects and items on homesites is helpful for understanding the analogy of virtual object and virtual fashion.

Much analysis has been done on person-object relations. Objects people use may represent the relation of one to oneself, to one's fellows, and to the universe (Csikszentimihalyi and Eugene 1981). A person's treasuring or forsaking an object may reveal his or her change. For instance, some object may be very treasured by somebody but may be forsaken later. This may reveal the owner's change of taste or may mean that

the object is no longer valuable for the owner as time goes by.

A homesite may also serve the purpose of representation. It may show the relation of a man to himself, to his friends, and to the world. The modification of a homesite may mean that something has changed in the owner's life. The change can be the owner's taste, interest, knowledge, experience, or relation to others.

Besides the change of the person himself, there are also changes prompt by the shifting meaning of objects. This is what fashion is about. For instance, if a homesite has not been changed for a long time, its content should be out of date since other homesites have been changing a lot.

With the interaction among people, some virtual objects become popular and spread throughout the Web. This diffusion constitutes virtual fashion. In the following sections, I review phenomenon of real-world fashion and apply similar ideas to the virtual world.

2.3 Fashion Cycle - From the Physical World to the Virtual World

The fashion cycle refers to the introduction, acceptance, and decline of a fashion. In the following section, historical observation on clothing fashion is compared with virtual fashion to show a possible future trend of the virtual fashion cycle.

The acceleration of fashion cycle

Even though fashion need not be about clothing, clothing fashion is the kind of fashion that has been studied for the longest time. With some observation and analysis on clothing fashion, we may get some insight into the mechanism behind fashion.

The change of clothing fashion and the variation of duration of fashion cycles have been continuing since the thirteenth century. Especially, the pace of fashion cycle has greatly

accelerated since the nineteenth century. Before that, it often took decades for one dress style to succeed another. But today a new style often lasts for no more than one or two seasons (Davis 1992). Does this phenomenon mean something to virtual fashion? We may look into this issue by examining the factors that contribute to the acceleration.

Factors that may contribute to the acceleration

Several reasons may contribute to the acceleration of fashion cycle. One important factor is the fluidity of information in a social hierarchy. In the middle age, there was little fashion because the royalty and the poor were so distinct that the poor were not allowed to learn the dressing style from the royalty. As time went by, the class boundaries became looser and thus the poor could dress in obsolete styles of the royalty. This information flow resulted in fashion, as the royalty created new fashion and the poor learned obsolete styles from them, and then the royalty created another fashion to display their social status. In the recent decades, the electronic media also greatly quickened the information flow (Davis 1992) and thus accelerated fashion cycle. Information flows no longer need to transfer physically from one person to another, but can diffuse instantly through various channels.

Will virtual fashion accelerate?

From the historical observation on clothing fashion, it is likely that virtual fashion will accelerate in the future. With the development of easier web-authoring tools, more and more people will be able to easily take part in the diffusion of virtual fashion. This is analogous to the expansion of the clothing market from the royalty to upper-middle-class women in the nineteenth. Also, since copying objects from a homepage may only take several clicks, adopting virtual fashion is much easier than buying a new clothe. In addition, as news services with data mining tools enable people to easier and faster know what is hot in cyberspace, this new type of media may also quicken the virtual fashion cycle as what traditional media did to clothing fashion.

More extensively, fashion should accelerate in a world in which information, as opposed to material goods and wealth, is the predominant cultural force. Changes in virtual fashion also serve as markers of people's access to information, since early adopters of fashion are likely to know what is the new thing. As discussed earlier, fashion occurs when there are information flows between class hierarchies or sub-cultures. If a class hierarchy is too rigid, or if there is no class difference at all, fashion is not likely to occur. Here is a big question: online, is there the social structure to support virtual fashion hierarchies? What is the structure like? Is it fragile or coherent? These are important research topics.

Well-known sites like Yahoo are in a similar role as what traditional mass media plays in the information flow. They highly quicken information flow and thus accelerate virtual fashion. In fact, everybody who owns a homesite may provide a new channel for the information flow, even though most of them are not as effective as famous sites.

In the future, people may observe the long-term changes of fashion cycles with data collected by my system and verify if the above inference is true - even though it may take many years to have a solid conclusion.

2.4 Theories of Fashion and Their Implication

Fashion in the real world is recurring. Although there are many different viewpoints regarding the nature and content of its cycles, some cautious generalizations seem to reasonably explain how fashion changes. The following sections briefly introduce the most important two sociological theories of fashion and discuss how they might be applied to the virtual world.

Trickle-down theory

Articulated by Simmel (1904), the trickle-down theory sees innovations as being adopted first by an elite, mostly to show the elite's superiority. Other groups subsequently adopt the innovation to establish their own superior status, with successive groups imitating their immediate superiors. The theory suggests that two conflicting principles act as driving force for fashion change. Following the principle of imitation, subordinate social groups seek to demonstrate their status by adopting superordinate groups' innovation. Following the principle of differentiation, superordinate groups respond by initiate change in fashions. Trying to hold their status markers and preserve the status difference, superordinate groups are urged by a bottom-up force when old status markers are obsolete, as shown in Figure 1.

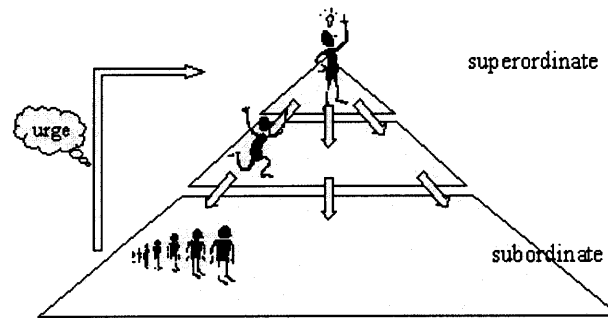


Figure 1: Trickle-down theory

This theory has several strengths. First, it gives us an understanding of how the social context in which fashion movement occurs will determine its direction, tempo, and dynamics (McCracken 1988). Second, it shows us that the two driving forces of fashion are mutually presupposing. Third, it may help give the fashion observer early warning of fashion change when there is a change in the behavior of an adjoining group.

However, clothing fashions in recent centuries are neither as universal nor as representatively focal as before, nor do fashions today seem able to enforce uniform-like compliance throughout a society (Bell 1947). Therefore, many scholars have been proposing different modification to the trickle-down theory to reflect the fashion pluralism and polycentrism (McCracken 1988). For example, Horowitz (1975) noted that mass fashion has significantly replaced elite fashion and resulted in the lack of social

hierarchy between individuals. King (1963) presented a "trickle-cross" model because media exposure allowed simultaneous adoption of new styles at all levels of society, instead of the trickle-down dispersion. Each level is led not by superordinate groups but by its own fashion innovators. Blumber (1969) argued that in a modern society, "collective selection" plays a much more important role than trickle-down effect, as will be discussed in the following section. These modified theories may reflect real as well as virtual world better than the original one.

Does trickle-down theory likely to work in the virtual world?

According to its hypothesis, the original trickle-down theory without modification does not work well in the virtual world, although it can still help us understand regional mechanisms. The revised versions of trickle-down theory may be closer to the mechanism of virtual fashion. There are several reasons for this.

First, it is likely that there are significant fashion pluralism and polycentrism in the virtual world. I want to use my system to verify this hypothesis. Nevertheless, the theory may still hold in some special cases, such as the upgrade of popular software announced by its official site.

Second, the class differentiation of homesites online is usually vague and fragile. Thus, the assumption of a classical hierarchy in society is not likely to hold on a large part of the Internet. As a result, the definition and complexity of classes needs to be refined for the virtual world. Researchers may want to verify if it is possible to observe various hierarchies, as suggested by the revised versions of trickle-down theory, or any multi-dimensional class hierarchy among homesites. To this end, a rough guess of adjoining groups may be based on link topology and the sequence in which homesites adopt virtual fashion.

Finally, so far it seems to be neither easy nor intuitive for superordinate group to observe if a fashion has been widespread in cyberspace. Consequently, the bottom-up force

resulting in superordinate group's differentiation might be difficult to perceive.

However, the revised versions of trickle down theory, which posit numerous and overlapping and even horizontal hierarchies, may provide important groundwork for modeling virtual fashion, as will be described later in the "Discussion on Theoretical Modeling" chapter.

Collective selection theory

Stated by Blumer (1969), the collective selection theory is also famous for conceptualizing the fashion process. Incorporating many ideas from collective behavior, the theory denies that hierarchical class relations animate the fashion process. It sees fashion as the gradual formation and refinement of collective tastes, which occur through social interaction among people with similar interests and social experience, with the result that many people develop tastes in common, as shown in Figure 2. Factors such as the historical continuity of fashion change, in which new fashions evolve from those previously established by the society, and the influence of modernity, through which fashions constantly respond to and keep pace with change in the larger mass society, also shapes the process.



Figure 2: Collective selection theory

This theory seems to fit well for the Internet because homesites on the Internet are usually distributed without clear social hierarchy. Since the observer must await the convergence

of taste in a particular direction, the theory can hardly give advance warning, and is much vaguer regarding modeling than the trickle down theory. Nevertheless, according to the collective selection theory, fashions that are more pervasive, more up-to-date, and more consistent with the current sociological environment have more chance to be more widespread. This idea may be helpful for guessing the popularity change of objects. The role of my system can still be the agent showing the most up-to-date emergence of popular objects.

Real-world fashion phenomena are difficult to track, but virtual fashion is quantifiable. In order to observe virtual fashion and see how the above-mentioned phenomena occur online, I have developed a system to measure virtual fashion, as will be introduced in the following chapter.

3. System Design

3.1 System Overview

My goal is to follow the spread of fashion online over time and analyze it. My initial approach is to examine a set of selected homesites each week and track the spread of links. The reason the system tracks only links is that the spread of links is easier to track precisely than images and other objects, and archiving only links diminishes the concern about privacy invasion.

As shown in Figure 3, the system starts from traversing and archiving a set of homesites. From the archive, the extracting subsystem extracts popular links. Since link is a kind of virtual object, for extensibility, link is also called "object" in this paper. The extracted popular objects are the basis of both the clustering subsystem and the integrating subsystem. All popular objects are clustered by the clustering subsystem into groups and subgroups, according to how related they are. In addition, the integrating subsystem accumulates the changes between adjacent weeks and shows the long-term popularity change of objects.

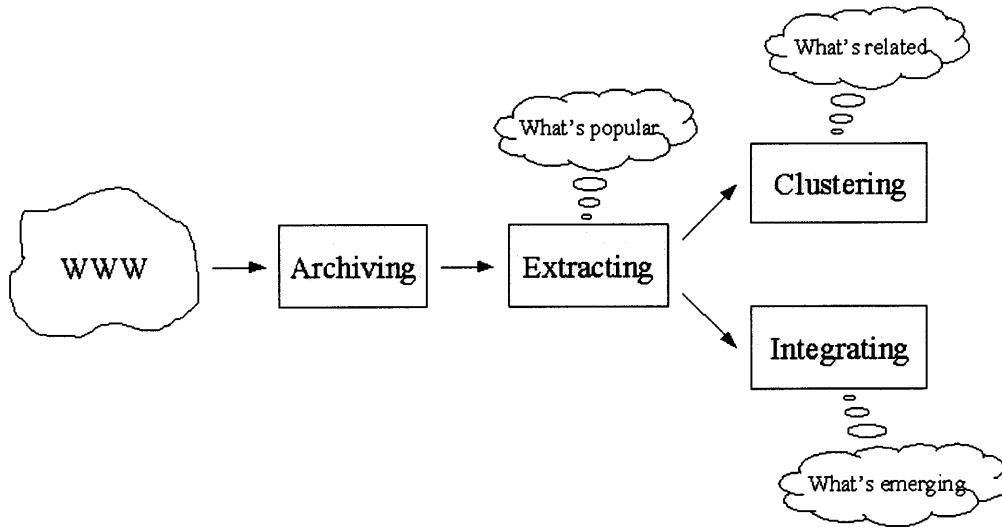


Figure 3: System diagram for virtual fashion project. Starting from archiving, the system extracts popular objects, clusters them, and integrates the data to show the long-term popularity change.

The extracting subsystem finds out what is popular in the domain, the clustering subsystem shows how things are related, and the integrating subsystem presents emerging virtual fashion.

The following sections introduce the subsystems and point out challenges and my experience in developing them.

3.2 Archiving Subsystem

Before the system can analyze virtual fashion, it needs an archive of homesites. This is made by the archiving subsystem, as shown in Figure 4, with its threefold purpose. First, it identifies all pages to traverse by following the branching network of hyperlinks, beginning with a set of initial pages that I have assigned. Second, it retrieves, filters, and stores those pages locally. Third, it tracks exactly the same set of homesites every week. This periodic nature distinguishes the archiving subsystem from traditional web robots. This section describes the initial set of homesites I chose to archive and relates the issues and challenges encountered in developing this specially designed archiving subsystem.

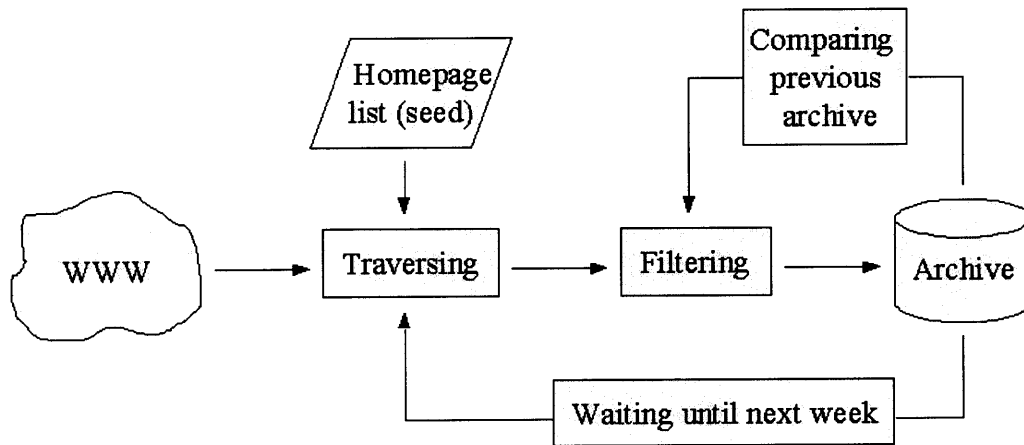


Figure 4: The archiving subsystem. It traverses through selected homesites, filters out useless or redundant data, and saves useful information into the archive.

The range of sample homesites being archived and why I choose them

I am particularly interested in looking at fashion as it is manifested on the homesites of individuals, instead of homesites of corporations or products, because of the role fashion plays in creating a presentation of self.

To begin with, I chose to archive four categories of homesites since February 1999:

- (1) 3000 homesites chosen from the Heartland community on GeoCities, which consists of homesites of people concerned with "parenting and hometown values."
- (2) 3000 homesites chosen from the Area51 community on GeoCities, which consists of homesites of people interested in paranormal phenomena.
- (3) The MIT Media Lab personal homesites, which consist of about 200 academically, oriented homesites.
- (4) The Yahoo list of people whose initial is A, which consist of about 2000 homesites all over the world.

I chose GeoCities because it is set up for personal homesites with lots of community-building mechanisms. I chose two different areas on GeoCities in order to see if different people with different interest have different patterns of social interaction. I chose the Media Lab homesites because they are what I am familiar with. I can compare the data with real-world events and see if the data make sense. I chose Yahoo list of people whose initial is A in order to get a random sample of homesites for comparison. These four categories contain different types of homesites. In this way, I may observe different types of ecology in different virtual communities.

Challenges

Besides the traditional functionality such as opening hundreds of concurrent connections and efficiently traversing through various web servers, the requirement of precisely archiving the same set of homesites periodically results in many more challenges as well as advantages over traditional robots, as described in the following subsections.

Reconfirmation

The archiving subsystem needs to reconfirm inaccessible pages, in contrast to traditional robots' feasibility of ignoring them. Traversing without this reconfirmation would result in the illusion of changes of the homesites due to temporary network errors. When receiving an error message, the subsystem should identify if the error may be temporary. If it may be temporary, the subsystem will try to retrieve the page again later for several rounds. My experience shows that archiving without this confirmation results in a lot of noise of the archive due to temporary errors. These errors sometimes result from my opening too many concurrent connections to a personal web server; sometimes they are because the server or its network is down for a while; sometimes they are just because too many people are also connecting to the server so it cannot afford any more connection. Numerous possibilities may cause temporary inaccessibility of a page, which doesn't matter to a traditional robot, but may make the archive unfavorable when being compared

to data of other weeks, as will be discussed in later sections.

For maximum efficiency, errors such as "HTTP 404 Not Found" may be ignored, while other errors such as "HTTP 408 Request Timeout," "Network Unreachable" or "Unknown Socket Exception" should be reconfirmed because they are very possible to be temporary.

Reusing previous archive to save space

The subsystem should take advantage of data of previous weeks to save both my resources and web servers' resources. For example, it can save a lot of time and storage by sending the If-Modified-Since header field of HTTP and checking the Last-Modified entity-header field of a page on an HTTP 1.0 compatible web server (Berners-Lee et al. 1996). When visiting an HTTP 1.1 compatible server, the performance may even be improved by taking into account HTTP 1.1's new caching mechanism, such as the opaque validators (Fielding et al. 1999). However, many servers still have not supported these headers, thus more space can be saved by checking to see if the page is the same as that of the previous week, stored locally with mechanisms such as the MD5 checksum (Rivest 1992).

Other required special features

The archiving subsystem filters all incoming data and saves only useful information. It tries to discard advertising, such as banner bars, and unrelated files based on certain heuristics. This not only saves space but also reduces meaningless dynamics/noise of the data. Otherwise it would result in much fake dynamics of pages. To filter advertising on a global scale, the subsystem may have the incoming pages filtered by a special-purpose proxy server dedicated to filtering advertising on Web pages (Siemens 1999, Brightwater 1999, ADscience 1999).

The archiving subsystem also records the last-modified time of each page for future reference. If it has not been modified since the last round, a page is saved as a pointer to

its last record in my database.

Many people put a redirecting HTML document as the root of a homepage. This document redirects readers to the actual URL of the homepage. A robot that aims at grabbing a complete homepage may also take into account this possibility and visits the actual homepage without being stopped by the different domain.

3.3 Extracting Subsystem

This subsystem extracts popular links from the archive of homepages. As shown in Figure 5, it starts by traversing all the HTML documents in the archive, parses them, converts relative URLs to absolute ones, and then creates the index of popular objects and homepages referring to those objects. In this way, it discovers popular objects, which are linked by more than one homepage.

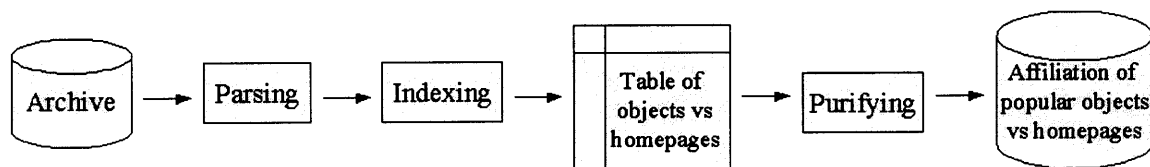


Figure 5: The extracting subsystem. It parses the archive, indexes and finds popular links, purifies the list, and then creates the database of popular links and homepages referring to them.

Refining the index

The index needs to be refined in several ways. First, more than one page in a single homepage may link to the same object. These pages should be grouped together under the name of the root URL of the homepage to avoid duplicate counting of homepages linking to the object. Second, links from CGI output should be filtered out because many of them are banner bars whose content change automatically and result in a lot of manipulation of links that are not actually done by people. The extracting subsystem is extensible so that

other modules for handling images, sounds, texts, and even Java Scripts may be added in the future.

Increasing the efficiency

For maximum efficiency, the extracting subsystem may be combined with the archiving subsystem so that the system can parse, index, and archive homesites at once, without parsing the HTML documents twice. Given enough memory, the system does not need to maintain an archive of the set of homesite, but only needs to keep the final index of objects with URLs of homesites referring to them. In my case, the original data on the set of eight thousand homesites exceeds 1GB, the filtered archive without compression is about 300 MB every week, and the weekly list of popular objects versus homesites is only several MB. All the subsequent subsystems require only this list but not the archive.

The extracted virtual fashion in the form of popular links are automatically updated and shown as a list with referring homesites on the project web site:

<http://vfashion.media.mit.edu>

3.4 Integrating Subsystem

Comparing the data of adjacent weeks can show the popularity change of objects. Each week, a list of newly linked and unlinked objects and their corresponding homesites are generated. By accumulating the changes over time, the integrating subsystem may help reveal the trend of the popularity of online objects. Long-term observation, which is automatically, updated on the project web site, on these changes helps show how virtual fashion evolves. Examples are shown later in the "Output and Case Studies" chapter.

The following are some finer details about the techniques for developing the integrating subsystem.

Smoothing function

The accumulation of change of popular objects also generates noise that results from temporary network errors and other reasons. Therefore, developing heuristics for eliminating noise is important to this subsystem. A smoothing function can be applied to diminish noise in the temporal data resulted from the unreliability of network or web servers. That is, some homesites are unreachable in one week and reachable in another, which result in spurious dynamics. The noise can be detected by looking into the data of adjacent weeks. The idea is that if a page is only unreachable in a certain week but remains exactly the same in other weeks, its absence in the single week can be ignored. In brief, while I do want to highlight step functions (item was present for a while, then gone for a while or vice versa), I also want to eliminate spikes (item present for a while, gone briefly, then present again).

Taking advantage of the periodic nature to save time

In the extracting subsystem and integrating system, a huge amount of data needs to be analyzed again and again every week. Continuing the analysis from the result of the previous week can save much time. The actual approach is dependent on each individual implementation and should be carefully designed to avoid loss of accuracy.

Other analysis and Meta data

Other analysis can also be done on the temporal database. For the clustering algorithm, which will be, mentioned later, information about the total number of links and popular objects each homesite links to is counted for future reference. Other statistics of the dynamics of homesites over time, people's tendency to update their pages, objects that show similar temporal change of popularity, and so on, can also be extracted from this temporal database. This statistics are valuable reference for researchers.

3.5 Clustering Subsystem

The relation between individuals as well as virtual objects is multidimensional. To make the relation clear and infer communities on the Web, I develop a clustering algorithm that delineates the sub-culture hierarchies based on how individuals get involved in the dispersion of virtual fashion. To this end, a relatedness function is chosen to satisfy a set of mathematical conditions. The conditions are deduced from how people may share common interests through placing common objects on their homesites.

Following I will first briefly conclude previous research that provides important background, and then show the intuitions which infer the mathematical conditions of my algorithm. A mathematical function that satisfies the conditions will be chosen to infer the relatedness between homesites. Based on the function, I use the hierarchical clustering algorithm to cluster homesites.

Inferring communities from link information

As the WWW grows in size and complexity, inferring high-level structure on the Web becomes increasingly important. There has been a growing amount of work (Botafogo et al. 1992, Pirolli et al. 1996) directed at the integration of textual content and link information to infer structures of communities on the Web.

It has been found that co-citation analysis (Small 1973) can be helpful for identifying interesting clusters of pages on the web (Pirolli and Pitkow 1997). The co-citation analysis has been very helpful for categorizing scientific papers according to how articles cite one another. In fact, both studies in co-citation and bibliographic coupling (Kessler 1963) can be inspiring to infer relatedness between pages. For two documents p and q , the co-citation quantity is equal to the number of documents cited by both p and q , and the bibliographic coupling quantity is the number of documents that cite both p and q . The larger these quantities are, the more likely p and q are about research in the same field.

Donath also use a similar concept to develop the "Visual Who" system (Donath 1995b) to visualize the social and organizational structure within an electronic community.

Interesting communities of pages can also be found on the Web through an analysis of link topology (Gibson et al. 1998). The communities can be viewed as containing a core of central, "authoritative pages" linked together by "hub pages"; and they exhibit a natural type of hierarchical topic generalization that can be inferred directly from the pattern of linkage.

A set of closely connected sites can also be grouped together by the clan graph (Terveen and Hill 1998) from a set of seed documents based on connection between them. The clan graphs treat links as undirected.

My approach - developing a relatedness function based on cultural dispersion

In this section, I present my method for inferring communities on the Web. It delineates the sub-culture hierarchies based on how individuals get involved in the dispersion of online objects. To this end, a relatedness function is derived from a set of mathematical conditions. The relatedness function can infer much more detailed relatedness between homesites than previous methods, which use simple relation, either binary (connected or unconnected) (Terveen and Hill 1998) or use the sum of common links (Pirolli and Pitkow 1997). It helps to generate meaningful group hierarchy even if the node-to-node distance of pages is not taken into account.

Moreover, the method differentiate homesites from objects on them, instead of mixing them all in a hypertext structure (Gibson et al. 1998, Terveen and Hill 1998), so that it can take into account virtual objects, whether they be HTML documents or not.

Finally, the relatedness function can work well independently of whether a page or a site is chosen to be the basic unit, so that the program may discover communities of documents (Pirolli and Pitkow 1997, Gibson et al. 1998) as well as communities of people. In the context of virtual fashion, the communities of people mean a lot more than communities of documents in that they may indicate further areas of research about particular sociological aspects of the Web.

The basic hypothesis

Inspired by research on co-citation analysis (Small 1973), we can begin to deduce the relatedness of homesites from the number of common objects, such as links to the same set of URLs. In this paper, I will use hypertext links as the objects we refer to for the purpose of simplification.

The intuition is that if two homesites both link to the same set of URLs, their owners may share similar interests and thus may be involved in the same sub-culture. The more links they have in common, the more likely the owners share similar interests.

In other words, I start with counting the co-citation quality of two homesites, but not the bibliographic coupling quantity (Kessler 1963). This is because the bibliographic coupling quantity is difficult to calculate precisely unless having a complete set of the WWW, and it cannot show the shared interests of the authors.

Note that I regard a homesite as the online identity of its owner. Thus, “homesite A and B may share the same interests” means that their owners may share the same interests.

Mathematical conditions

First, a function for quantitatively inferring the relatedness between any two homesites is needed. This should not be a binary (connected/unconnected) relation, because a binary value discards too much useful information about different degrees of relatedness.

The simplest function one may come up with is to calculate the total number of common links between two homesites, as used in traditional co-citation analysis (Small 1973) or Pirolli and Pitkow 1997. Although it works well for bibliography, this method is inaccurate for analyzing the Web since homesites vary significantly in size. This is resulted from the fact that publishing hundreds of papers is quite difficult, but putting hundreds of links on a homesite is relatively easy. In other words, the number of references of a paper is usually below one hundred, but the number of links on a web site is usually much more than that. Thus, by counting only the total number of common links, portal sites may be highly related to most homesites. It is because portal sites contain so many links that they tend to have links in common with any other sites, while in fact portals are not especially affiliated with most of these homesites.

On the basis of this observation, one may refine the function to divide the number of common links by the number of total links on each homesite. However, this method results in very distorted and unfavorable measurements: the denominator (the number of total links) has such a great effect that the numerator (the number of common links) rarely makes a difference. The size of the homesite turns out to be the dominant, and almost the only, factor.

To derive an appropriate function, a list of mathematical conditions should be helpful. To begin with, let us think of the affiliation of a homesite A to any URL it links to. If A links to t_A URLs in total, it is intuitive that the smaller t_A is, the higher the affiliation A may have to any URL it links to. For example, if both homesites A and B link to

http://www.media.mit.edu but A links to a total of 10 URLs and B links to a total of 1000 URLs, it is likely that A is more affiliated with http://www.media.mit.edu than B.

Based on the idea of affiliation, we can think of the relatedness between homesite A and homesite B. Again, assume that A links to t_A URLs in total, and B links to t_B URLs in total, and they have c_{AB} links in common (an example is given in Figure 6.) It is obvious that A and B are more likely to share the same interests when c_{AB} becomes larger or when t_A and t_B get smaller.

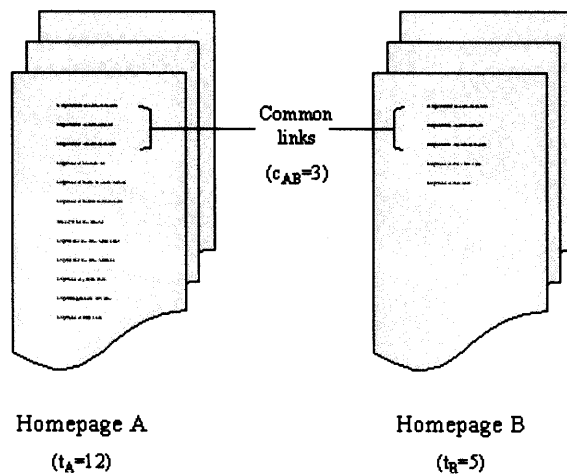


Figure 6: An example of having common links between homesite A and homesite B.

To simplify the problem, we may list the mathematical conditions for calculating the given homesite A's affiliation to the number of common links, c_{AB} . Later, by combining both A's and B's affiliation to the set of common links, we can get the final relatedness function between A and B.

As listed below, I have integrated the mathematical conditions for selecting the affiliation function which calculates A's affiliation to the links A and B have in common. Let's call the affiliation function $f(c_{AB}, t_A)$. The first parameter is the number of the links, which are linked by both A and B, we want to count homesite A's affiliation to. The second

parameter is the total number of links on homesite A. The parameters are always natural numbers.

$$(1) f(n, m) < f(n+1, m), \text{ where } n+1 < m.$$

The larger the number of common links, the greater A's affiliation to this set of links. For example, two out of three links on A implies more affiliation than one out of three links on A.

The concept is the same as co-citation analysis.

$$(2) f(n, m+1) < f(n, m), \text{ where } n < m.$$

The larger the number of total links on A, the less A's affiliation to a certain set of links on it.

For example, one out of three links on A implies more affiliation than one out of one hundred links on A.

The concept is the same as the refinement mentioned in the previous section.

$$(3) f(n, m) < f(an, am), \text{ where } n < m, \text{ and } a > 1.$$

My experience made me think that even though the fraction n/m is equal to an/am , the larger the numbers, and the more concrete evidence of affiliation. This empirical condition may not be strong enough all the time. Should it not be helpful, we can get rid of the coefficients k_m and k_n in the relatedness function mentioned later.

For example, three out of nine links on A implies more affiliation than one out of three links on A.

(4) $f(n, m) < a f(n, am)$, where $n < am$, and $a > 1$.

From experience, the decline of the value of affiliation should not be too reactive to the total number of links, as the lower curve on Figure 7. Otherwise, the total number of links will be the only dominant factor of the function. A less steep curve, the upper curve on Figure 7, may resolve this problem. Thus, $f(n, am)$ should be larger than $(1/a)$ times $f(n, m)$.

This resolves the problem brought by the refinement mentioned in the previous section.

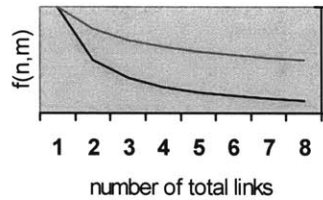


Figure 7: Possible change of $f(n,m)$ when $n=1$. The lower line represents a function that reacts too much to m , the number of links on the homesite in total.

Relatedness function

There can be many possible functions to satisfy these mathematical conditions. The affiliation function I chose is:

If $n=0$,

$$f(n, m)=0$$

else

$$f(n, m) = \frac{\frac{1}{n^{r_n}} + k_n}{\frac{1}{m^{r_m}} + k_m}, \text{ where } 0 < k_n \leq k_m, 1 \leq r_n < r_m.$$

r_n is the order of the root taken of n , r_m is the order of root taken of m .

The reason that r_m should be larger than r_n is that m is always larger and varies a lot more than n . To avoid m becoming the dominant factor, r_m needs to be larger than r_n . The coefficients: k_n, k_m, r_n, r_m can be determined by the feedback from the actual data. Experiments show that to set r_n around 1 and $r_m \geq 2$ works well.

If you want to make certain mathematical conditions play a more important role, you may adjust the coefficients. Even though changing the coefficients will affect the importance of each mathematical condition, all the four mathematical conditions will be satisfied regardless of the coefficients chosen. Examples of refining the coefficients are shown in Figure 8 and Figure 9.

Thus, the relatedness function between A and B can be labeled $g(f_A, f_B)$. The related function can then be set as, for example,

$$g(f_A, f_B) = \sqrt{f_A f_B},$$

which is the geometric mean of A's and B's affiliation to their common links. This function is one of the simplest functions that fit the mathematical conditions I specified. As long as a function fits the conditions and has appropriate coefficients, I suppose it to behave reasonably well. Note that even though better equations that satisfy the mathematical conditions may be found and analyzed by a series of simulation and human rating, finding the most optimized equation that satisfies the four conditions is beyond the scope of this research. The most optimized equation may also differ according to the nature of communities and virtual objects being analyzed.

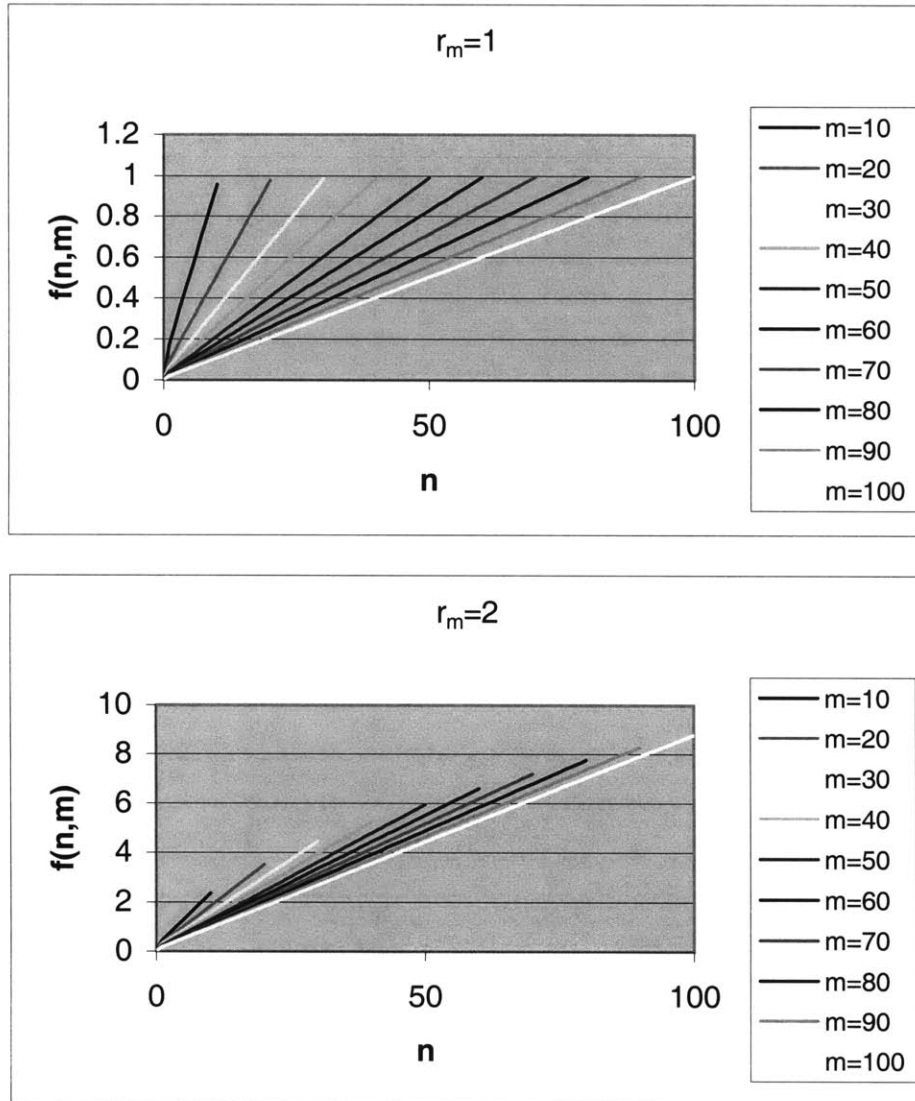
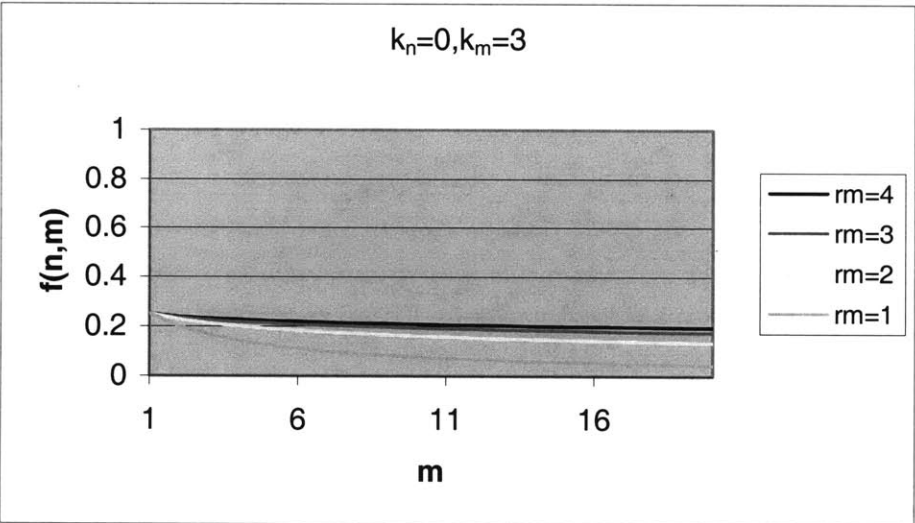
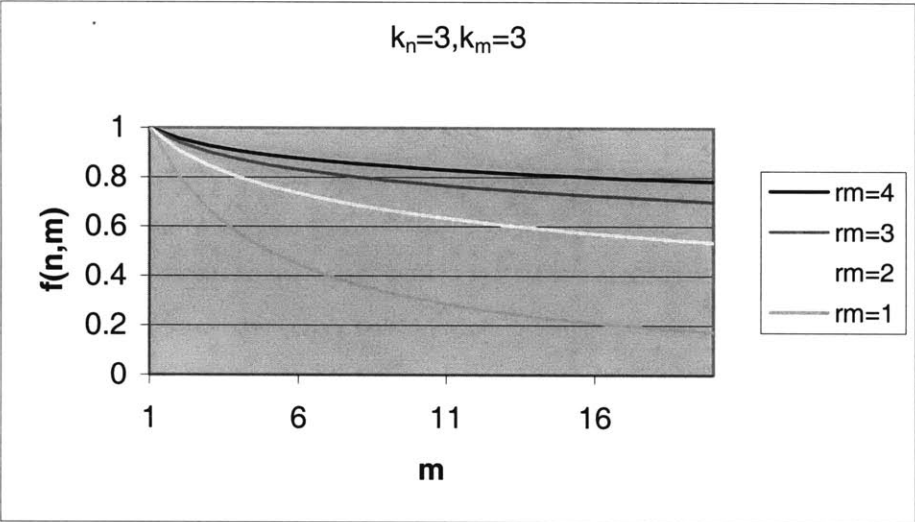
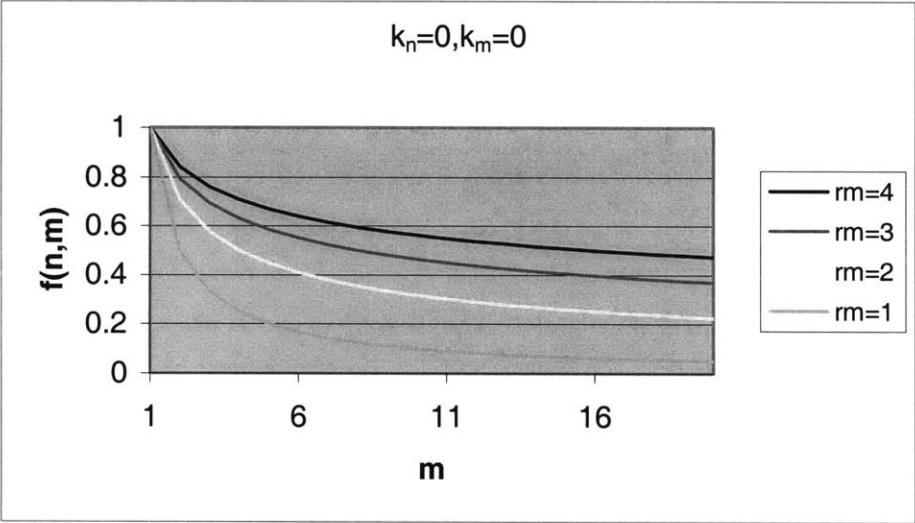


Figure 8: The two charts show samples of the behavior of the affiliation function for $r_m=1$ and $r_m=2$, respectively. The lower chart shows a more favorable choice of r_m .



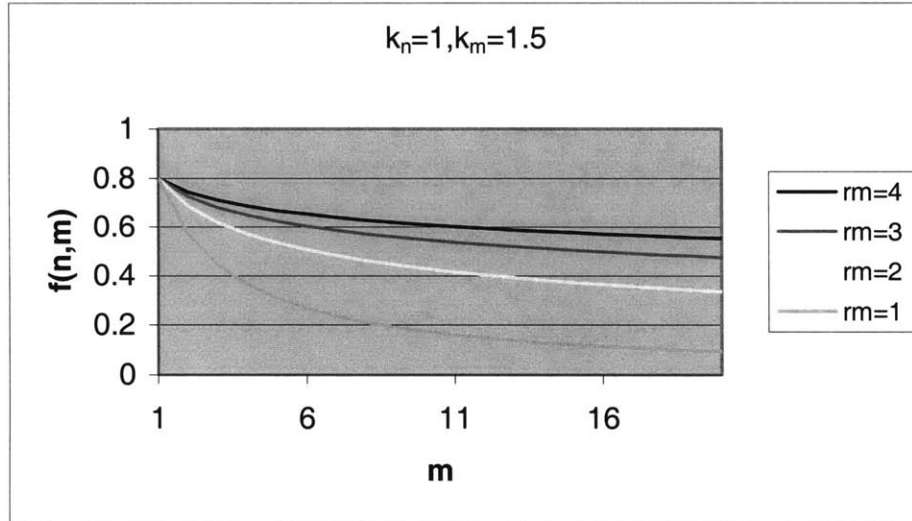


Figure 9: For sample case of $n=1$, above charts show the behavior of the affiliation function for different k_n and k_m . The bottom one is more favorable than others.

In the virtual fashion project, we refine the m to be the square root of "the number of total links on the homesites \times the number of links that are on more than one homesites." This makes the choice of m even more reasonable.

Discrimination value

For better accuracy, I also refine the function in the program based on additional criteria. For instance, instead of treating all links on homesites equally, weighting links that are shared by only a few people more can be more accurate. Links that are linked by only a few people may reveal the specific interests of their users better than those widespread links. For example, almost all homesite in the GeoCities have a link to the GeoCities front page, and only a few homesites have a link to a certain site outside GeoCities. Obviously, the latter reveals a lot more about a person's specific interest than the former. In this way, we may reveal sub-culture relations more accurately.

Besides, incorporating time as a factor may make the underlying information flow even clearer. Since a person may change his interest, the link a person recently links to can reveal more about his recent interest. The length of the period over which an individual

adopts a certain link may also be meaningful: short-period means short-term interest. Furthermore, some links may get popular in a short period of time. These recently fashionable links may have significant cultural meanings so that they get popular so fast. Therefore, the fashionable links can be weighted more than ordinary links. The statistics of the temporal dynamics can be obtained from other subsystems of the project.

Based on these observations, the relatedness function can be further refined to give every link different discrimination value. The actual implementation differs according to which criteria are more important for the specific application.

Clustering the homesites based on their relatedness

My goal is to reveal sub-culture hierarchies. Therefore, I cluster homesites based on the relatedness between them. By calculating the multidimensional relation between homesites with the relatedness function, my program categorizes homesites into hierarchical clusters (Everitt 1980). The relatedness is equal to the "distance" from the perspective of clustering algorithms. Starting from the highest relatedness observed, the program progressively decreases the threshold of relatedness to find homesites connected directly or indirectly to one another with relatedness above the threshold. Thus it gradually discovers sub-culture hierarchies within the set of homesites.

Using hierarchical clustering shows the hierarchical relation between groups and sub-groups, which is very helpful for our application. In this way, people with similar interest are in the same group. However, hierarchical clustering algorithm is not perfect. A drawback of the algorithm is that the structure of hierarchies sometimes changes a lot with only a minor difference in the data set. As a result, if you want to see the hierarchies of virtual fashions over time, the resulted hierarchies are not always smoothly transforming from one week to another. Anyway, if you want to see the hierarchical relations between groups and sub-groups, using other clustering algorithms (such as K-mean or Cluster-Weighted Modeling) may probably result in a similar problem because

you still need to recursively cluster those groups. Making the clusters continuous over time is an important and separate research topic, and is beyond the scope of this thesis. Some similar issues for the visualization system will also be discussed in the "Ideal and Reality" section in the "Output and Case Studies" chapter.

There are tons of issues regarding choosing an appropriate clustering algorithm, and there are lots of tricks regarding making the implementation efficient because it involves analyzing the multi-dimensional relation between thousands and thousands of items. For example, in the beginning, it took my program three days just to cluster a week of data on Heartland; with lots of algorithmic tricks, it takes less than a minute now. I will skip the details of these issues since they are off the focus of this thesis.

My approach generates meaningful group hierarchies even if the node-to-node distance of pages is not taken into account. Experiments with about eight thousand homesites show satisfactory results.

Discussion on the relatedness function

The relatedness function can infer much more detailed degree of relatedness between homesites than other contemporary methods, which use simple relation, either binary (connected or unconnected) (Terveen and Hill 1998) or use the sum of common links (Pirolli and Pitkow 1997). The relatedness function may also be incorporated into these contemporary approaches to refine their measure of similarity/relatedness between pages.

In addition to clustering homesites, the function can also be applied to classifying virtual objects. The hypothesis then becomes: if two objects are both linked by the same set of homesites, they may be related.

This method can also work well independently of whether a page or a site is chosen to be the basic unit. Regarding the diffusion of virtual fashion, the communities of people mean

a lot more than communities of documents. Therefore, this method fits well for analyzing the relationships between individuals and the diffusion of virtual fashion.

The system I developed have tracked and analyzed virtual fashion in the form of links on about eight thousand homesites for a year. The results and case studies are show in the following chapter.

4. Output and Case Studies

4.1 Online Service

To begin with, I provide a simple online interface, as shown in Figure 10. The URL of this service is <http://vfashion.media.mit.edu/vfashion/list.html>.

This service provides information about:

- (1) The weekly archive of popular objects in the form of links with their referring homesites;
- (2) Lists of popularity changes of links between adjacent weeks;
- (3) Lists of long-term observations on the popularity changes of links; and
- (4) Group hierarchies of links and homesites (from the clustering sub-system).

The data are collected from four sample areas on the Internet, as explained previously in the "archiving subsystem" section, including "Area51 on GeoCities," "Heartland on GeoCities," "MIT Media Lab homesites," and "Yahoo list of people whose initial is A."

Area 51 on GeoCities	Heartland on GeoCities	Media Lab	Yahoo Initial A
<u>2/28/1999</u> 2/28 - 3/6	<u>2/28/1999</u> 2/28 - 3/7	<u>2/28/1999</u> 2/28 - 3/6	<u>3/1/1999</u> 3/1 - 3/6
<u>3/6/1999</u> 3/6 - 3/13	<u>3/7/1999</u> 3/7 - 3/14	<u>3/6/1999</u> 3/6 - 3/13	<u>3/6/1999</u> 3/6 - 3/12
<u>3/13/1999</u> 3/13 - 3/22	<u>3/14/1999</u> 3/14 - 3/22	<u>3/13/1999</u> 3/13 - 3/21	<u>3/12/1999</u> 3/12 - 3/22
<u>3/22/1999</u> 3/22 - 3/27	<u>3/22/1999</u> 3/22 - 3/28	<u>3/21/1999</u> 3/21 - 3/27	<u>3/22/1999</u> 3/22 - 3/26
<u>3/27/1999</u> 3/27 - 4/4	<u>3/28/1999</u> 3/28 - 4/4	<u>3/27/1999</u> 3/27 - 4/3	<u>3/26/1999</u> 3/26 - 4/2
<u>4/4/1999</u> 4/4 - 4/12	<u>4/4/1999</u> 4/4 - 4/16	<u>4/3/1999</u> 4/3 - 4/12	<u>4/2/1999</u> 4/2 - 4/14
<u>4/12/1999</u>	<u>4/16/1999</u>	<u>4/12/1999</u>	<u>4/14/1999</u>

Figure 10: Virtual fashion archive homesite.

The weekly archive of popular objects

Figure 11 shows the list of popular objects (in the form of links) on March 21, 1999, in the MIT Media Lab domain. In the left frame, popular objects appear in the order of their popularity, i.e., how many homesites link to them, with the total number of homesites linking to them shown within square brackets to the left of the URL. Clicking on a number within square brackets brings out a list of the referring homesites in the right frame (Figure 12). Multiple referring pages within a homesite are put together so that the right frame displays how many people, instead of how many documents, link to the URL. Clicking on any URL brings out the content of the URL in the right frame (Figure 13).

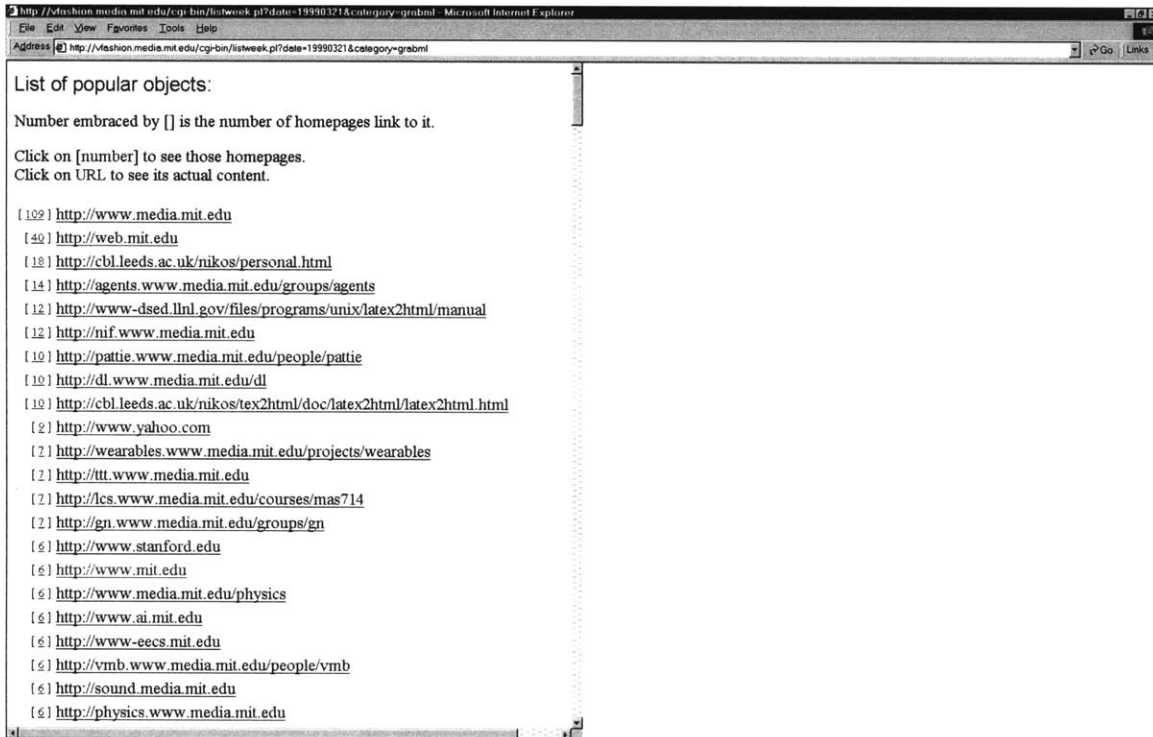


Figure 11: A list of popular objects on 3/21/1999 in the MIT Media Lab domain.

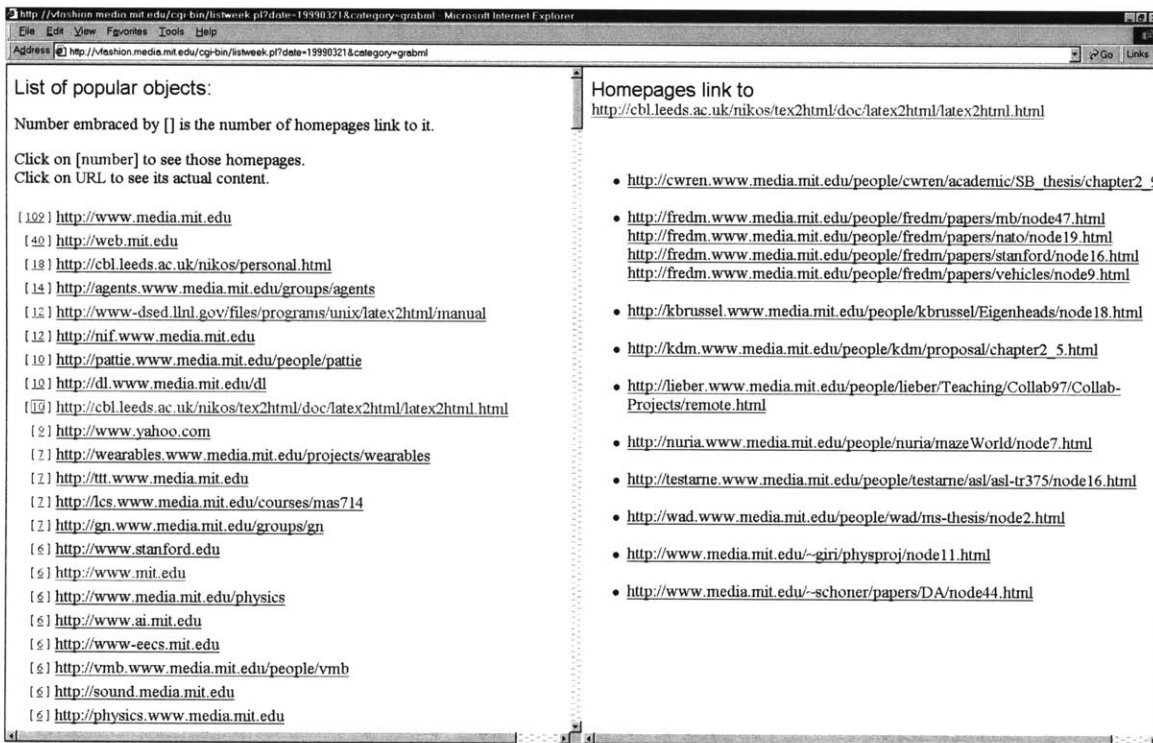


Figure 12: Clicking on the number within square brackets which is in front of an object brings out a list of the referring homesites in the right frame.

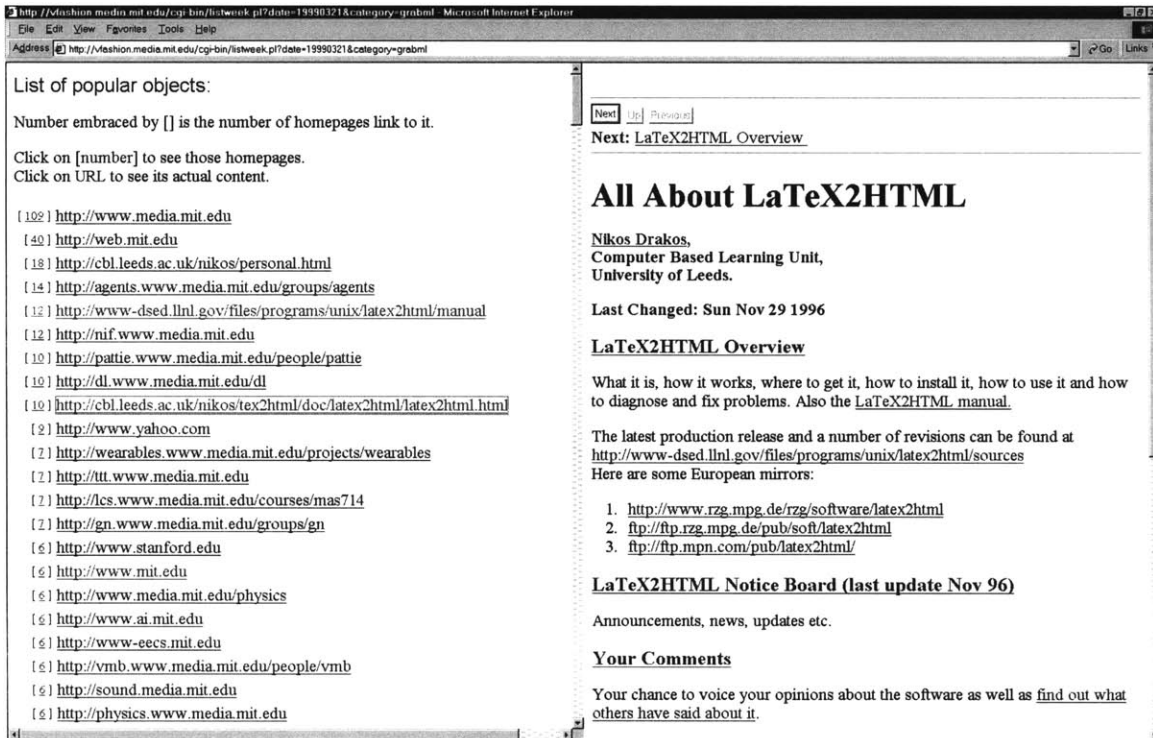


Figure 13: Clicking on any URL brings out the content of the URL in the right frame.

Lists of popularity changes of links between adjacent weeks

By comparing data of adjacent weeks, the system finds out what objects' popularity has been changed. The example in Figure 14 shows the popularity change of links between March 21, 1999 and the previous week. In this example, the popularity of seven objects in the Media Lab domain has changed during the week.

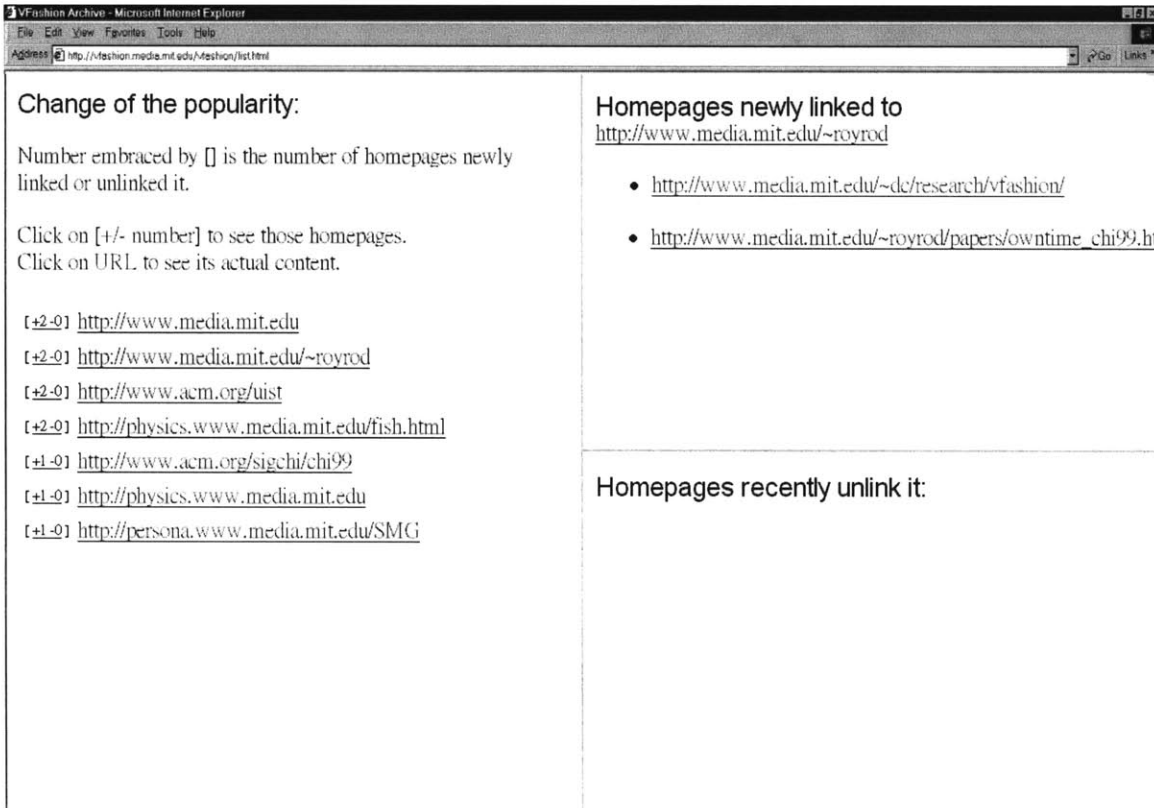


Figure 14: A list of objects whose popularity has changed during a week. The left frame lists the objects, the upper-right frame lists homesites newly linked to a chosen object, and the bottom-right frame shows homesites no longer linked to it.

Lists of long-term popularity changes

The system integrates the change of popularity of objects between weeks and shows the result on the online service as well (Figure 15.) Figure 16 shows the long-term popularity of objects, in the form of links, in the MIT Media Lab domain. The left frame lists popular objects. Clicking on the symbol [TIME] in front of an object brings out the details in the right frame (Figure 17). It shows the addition and subtraction of homesites that refer to the object every week. The numbers in parentheses represent the number of new referring pages that are part of homesites that have already linked to the object. In this case, the total number of homesites referring to the object will not increase because of the new pages. Finally, clicking on any URL brings out the content of the URL in the right frame (Figure 18).

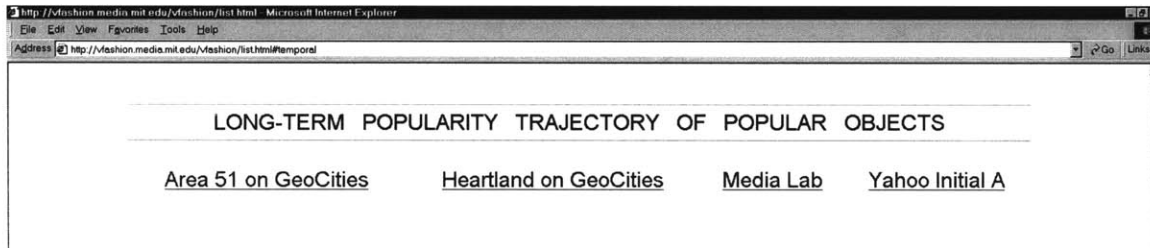


Figure 15: The online service that shows the long-term popularity change of objects.

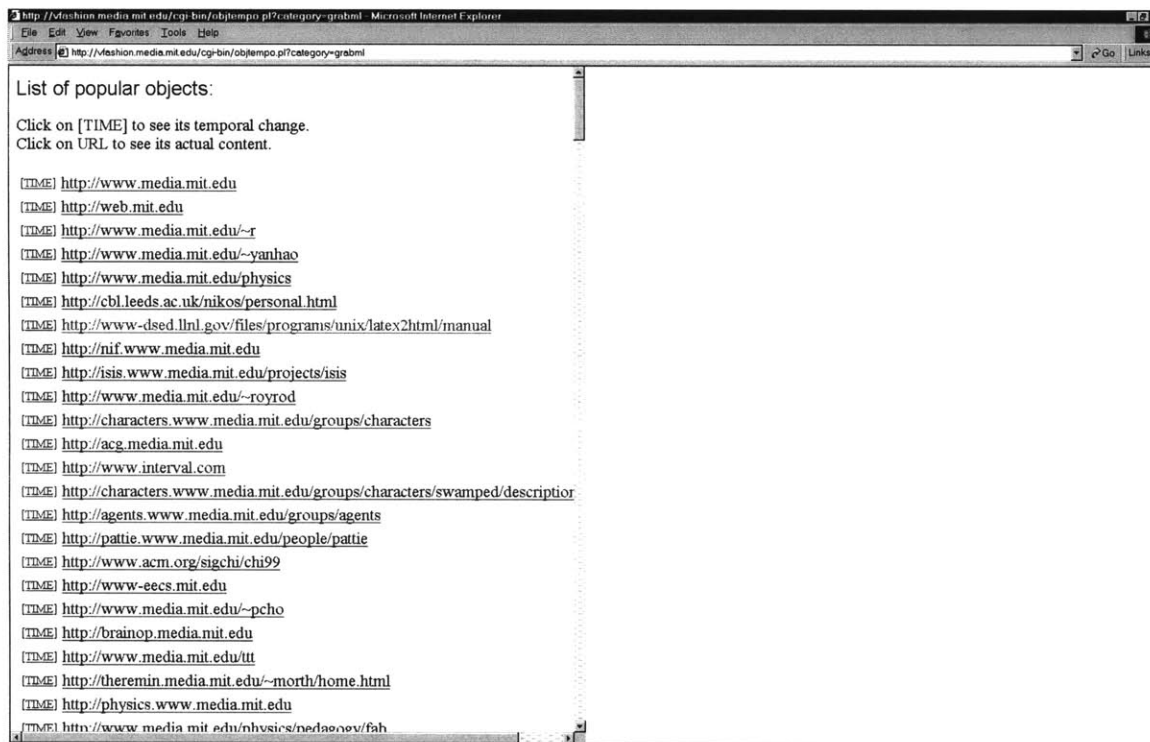


Figure 16: A list of popular objects in the MIT Media Lab domain.

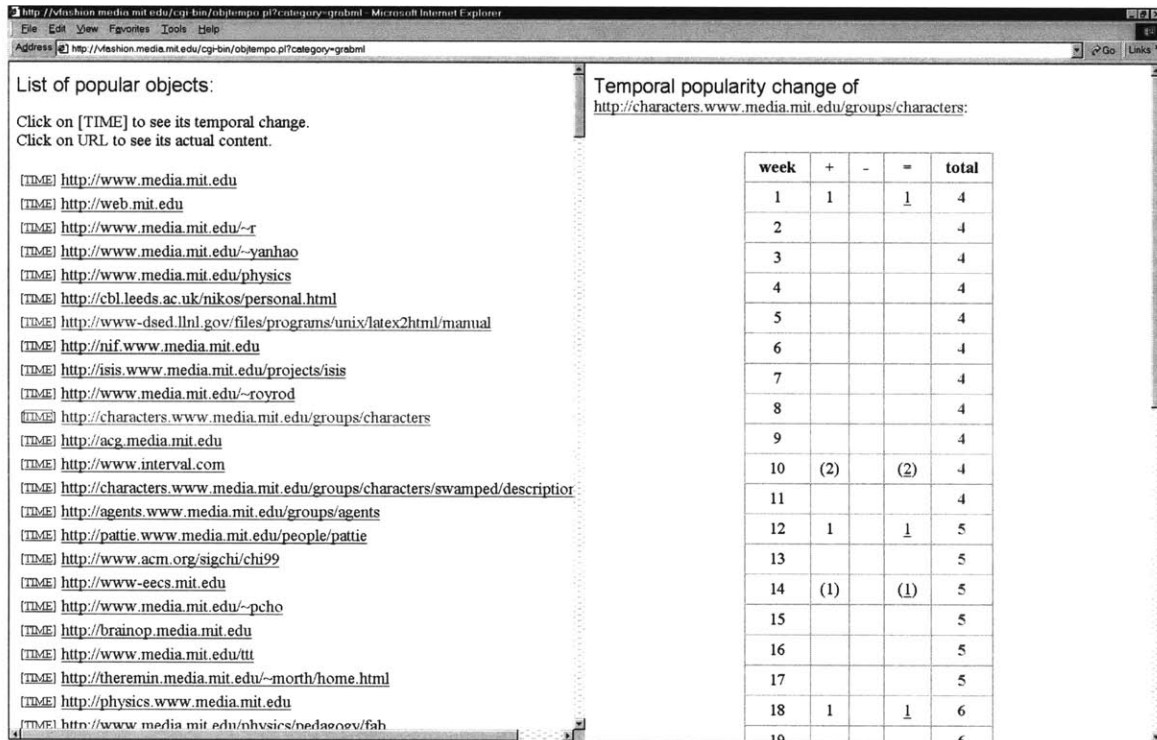


Figure 17: Clicking on [TIME] in front of an object brings out the details in the right frame.

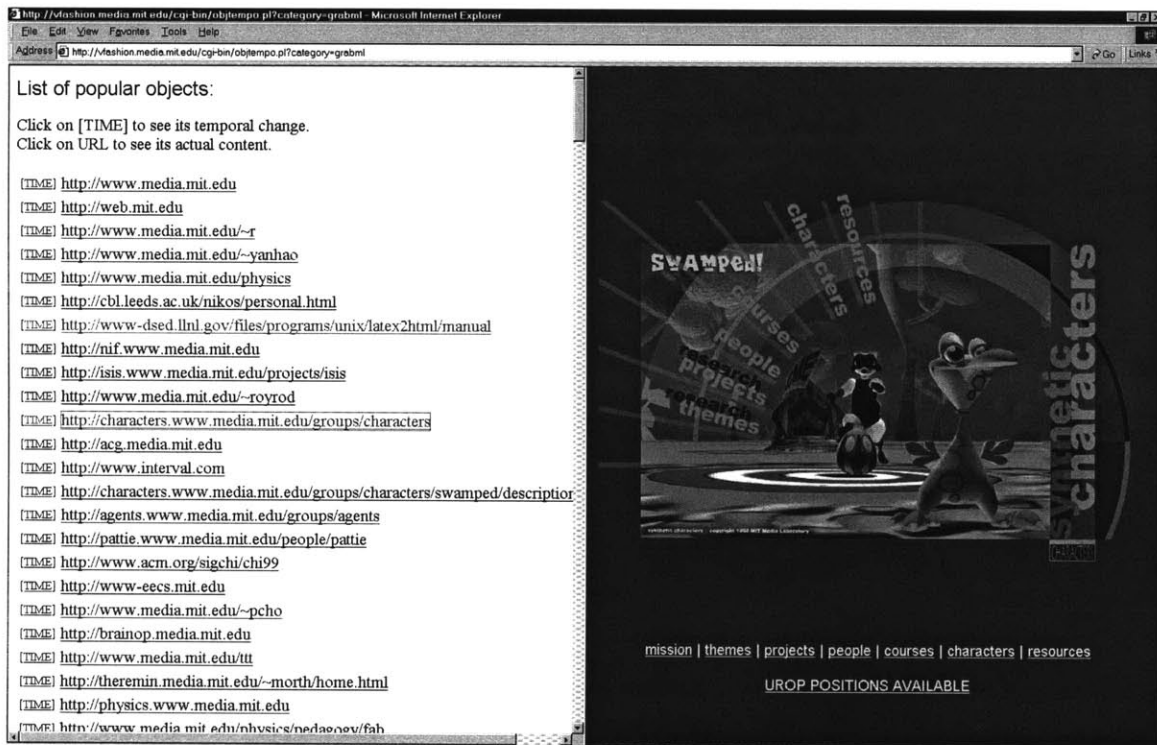


Figure 18: Clicking on any URL brings out the content of the URL in the right frame.

The data are important for studying sociology online. They not only show the ecology at a given time, but also reveal the temporal changes, which are important for studying social phenomena such as fashion.

The group hierarchy of links and homesites

The clustering subsystem classifies homesites and objects, respectively, into hierarchical clusters based on the relatedness between them. Figure 19 shows the list leading to the graph of these group hierarchies.

Figure 20 is a graph that shows the group hierarchy of homesites in the MIT Media Lab domain in a certain week. Each bar means a sub-group. Members in the same sub-group may share common interest. Bars connected together have the same parental group. The longer a bar is, the more related its members are. The number within parentheses indicates the number of member homesites in the subgroup. Clicking on a bar brings out homesites in the sub-group in the right frame, as Figure 21 shows. In this example, sandy and cwren are in the same sub-group and may share very similar interest.

http://fashion.media.mit.edu/fashion/list.html - Microsoft Internet Explorer

Address http://fashion.media.mit.edu/fashion/list.html#temporal

GROUP HIERARCHIES OF HOMEPAGES / OBJECTS

Area 51 on GeoCities	Heartland on GeoCities	Media Lab	Yahoo Initial A
2/28/1999: HP / OBJ	2/28/1999: HP / OBJ	2/28/1999: HP / OBJ	3/1/1999: HP / OBJ
3/6/1999: HP / OBJ	3/7/1999: HP / OBJ	3/6/1999: HP / OBJ	3/6/1999: HP / OBJ
3/13/1999: HP / OBJ	3/14/1999: HP / OBJ	3/13/1999: HP / OBJ	3/12/1999: HP / OBJ
3/22/1999: HP / OBJ	3/22/1999: HP / OBJ	3/21/1999: HP / OBJ	3/22/1999: HP / OBJ
3/27/1999: HP / OBJ	3/28/1999: HP / OBJ	3/27/1999: HP / OBJ	3/26/1999: HP / OBJ
4/4/1999: HP / OBJ	4/4/1999: HP / OBJ	4/3/1999: HP / OBJ	4/2/1999: HP / OBJ
4/12/1999: HP / OBJ	4/16/1999: HP / OBJ	4/12/1999: HP / OBJ	4/14/1999: HP / OBJ
4/17/1999: HP / OBJ	4/18/1999: HP / OBJ	4/17/1999: HP / OBJ	4/19/1999: HP / OBJ
4/25/1999: HP / OBJ	4/25/1999: HP / OBJ	4/25/1999: HP / OBJ	4/25/1999: HP / OBJ
5/1/1999: HP / OBJ	5/2/1999: HP / OBJ	5/1/1999: HP / OBJ	4/30/1999: HP / OBJ
5/8/1999: HP / OBJ	5/9/1999: HP / OBJ	5/8/1999: HP / OBJ	5/7/1999: HP / OBJ
5/15/1999: HP / OBJ	5/16/1999: HP / OBJ	5/15/1999: HP / OBJ	5/14/1999: HP / OBJ
5/22/1999: HP / OBJ	5/23/1999: HP / OBJ	5/22/1999: HP / OBJ	5/21/1999: HP / OBJ
5/29/1999: HP / OBJ	5/30/1999: HP / OBJ	5/29/1999: HP / OBJ	5/28/1999: HP / OBJ

Figure 19: The online service that shows the graphs of group hierarchy of homesites and objects.

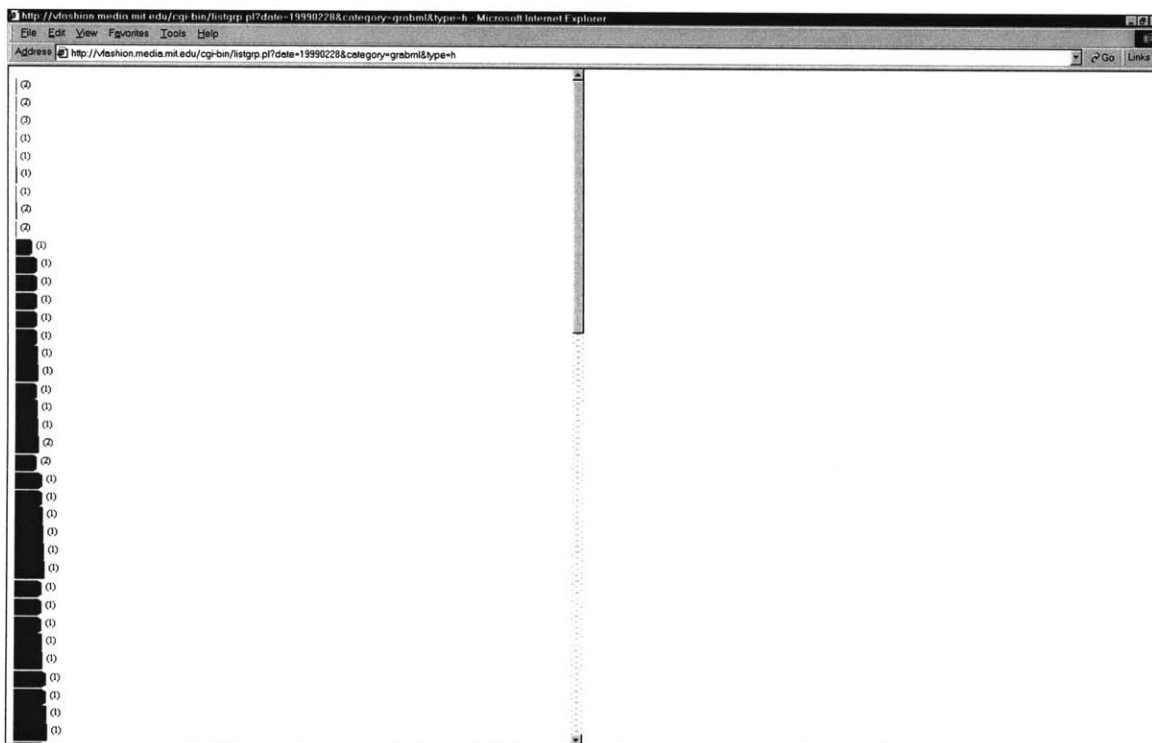


Figure 20: The graph that shows the group hierarchy of homesites. Each bar represents a subgroup. Bars connected together have the same parent group.

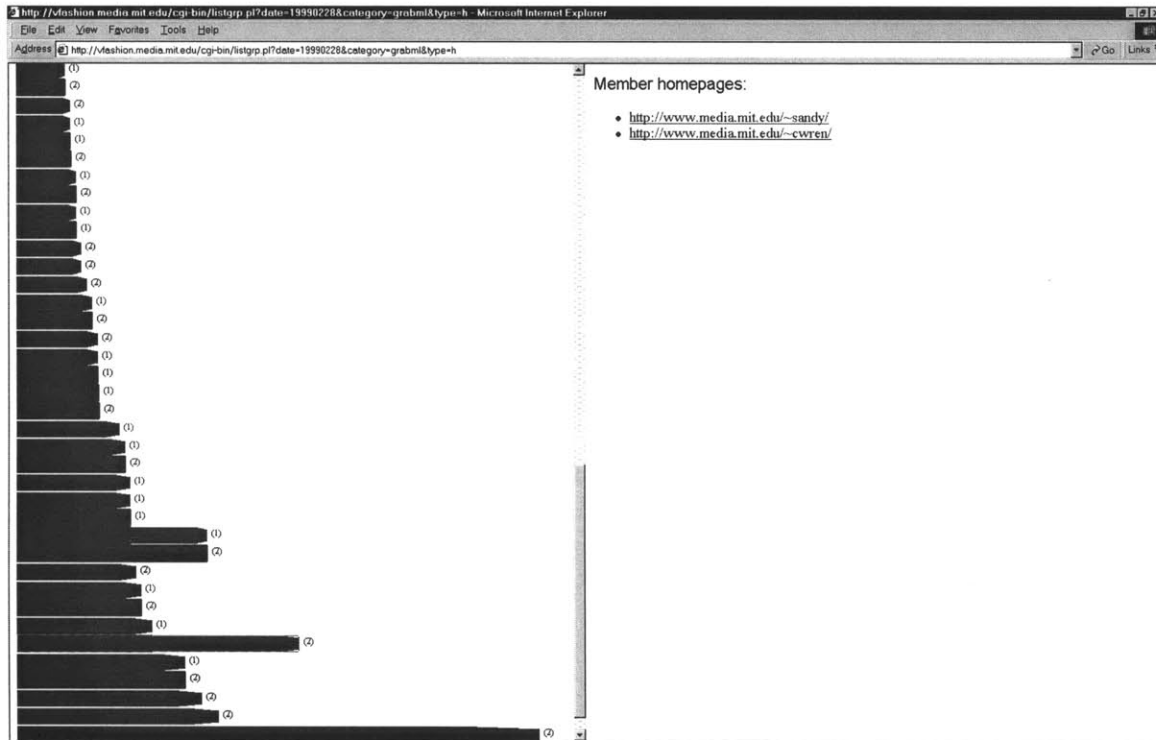


Figure 21: Clicking on a bar brings out member homesites of the sub-group in the right frame. The longer a bar is, the more related its members are.

4.2 General Findings

Different virtual communities

Figure 22 shows the related statistics in 41 weeks, from 2/28/1999 to 12/6/1999. The column titled "Min. Objects" shows the minimum number of objects in the domain in this period, while "Max. Objects" labels that of the maximum number. Note that in the "Yahoo People List With Initial A" domain, the number of objects is a lot more than that in other domains because there are various kinds of objects different people link to, but most of these objects are much less popular than objects in other domains. To the contrary, homesites in other more focused groups link to more common objects and thus the total number of common objects is lower. Besides, the column of "Changes in 41

weeks" includes the total number of changes, which means the accumulation of additions and subtractions of pages to popular objects.

Dynamics-A is the result of the equation:

$$\text{Dynamics-A} = (\text{Changes in 41 weeks}) / 41 / (\text{Average Number of Objects}).$$

The total change divided by 41 gives the average change every week. This value is then divided by the average number of objects to represent the density of the change.

Dynamics-B is the result of the equation:

$$\text{Dynamics-B} = (\text{Changes in 41 weeks}) / 41 / (\text{Number of Homesites in the Domain}).$$

The value of dynamics-A is more representative than the value of dynamics-B because the total number of homesites does not necessary mean the total number of active homesites. Many homesites, especially those on free web-hosting service sites, are dead. The way dynamics-A is calculated alleviates this problem even though the way dynamics-B is calculated may be more intuitive.

	Homesite Number	Min. Objects	Max. Objects	Changes in 41 weeks	Dynamics -A	Dynamics -B
Media Lab	159	1306	1523	1183	0.020	0.181
Area51	3000	8466	9475	25105	0.068	0.204
Heartland	3000	6961	9800	26629	0.078	0.216
Yahoo I.A.	1938	11665	21262	22260	0.033	0.280

Figure 22: Activities of popular objects in the four domains.

Obviously, much more virtual fashion activity occurs on GeoCities than among Media Lab homesites. People on GeoCities change their homesites more frequently and exhibit more coherent patterns of change. Within GeoCities, more virtual fashion activity and dynamics occur on the Heartland members' pages than on Area51. I hypothesize that those who use homesites to establish their online identities tend to get involved in the diffusion of online fashions. People on the family-themed Heartland may have more emotional contact and be more concerned with how others view their homesites. As a result, they are more involved in the fashion process than Media Lab researchers, who use homesites mostly as academic tools, and Area51 members, who may be more interested in science-fiction entertainment than in community-building.

Media Lab ecology

Based on the statistics of Media Lab homesites, we also found interesting phenomena that show how real-world activity affects the virtual world. Figure 23 shows an example. In February, the beginning of a semester, people updated homesites a lot. In fact, many of them linked to class homesites. As time went by, there were fewer and fewer virtual fashion activities, possibly because of the research load.

The fourth week was the week right before a Media Lab open house. For the open house, people worked very hard on their own projects and modified their web pages a lot. However, there was much less dynamics of virtual fashion activity in the week! It's the lowest point in the chart. In October when the other open house in the year was held, this happened again and resulted in the second lowest point in the chart. The October open house was longer and thus it diminished virtual fashion activity for two weeks. My hypothesis is that people were busy working on their own projects and homesites but did not have time to get involved in any diffusion of online culture.

Note that in the following figures, dates are labeled in the horizontal axis in the format of YYYYMMDD.

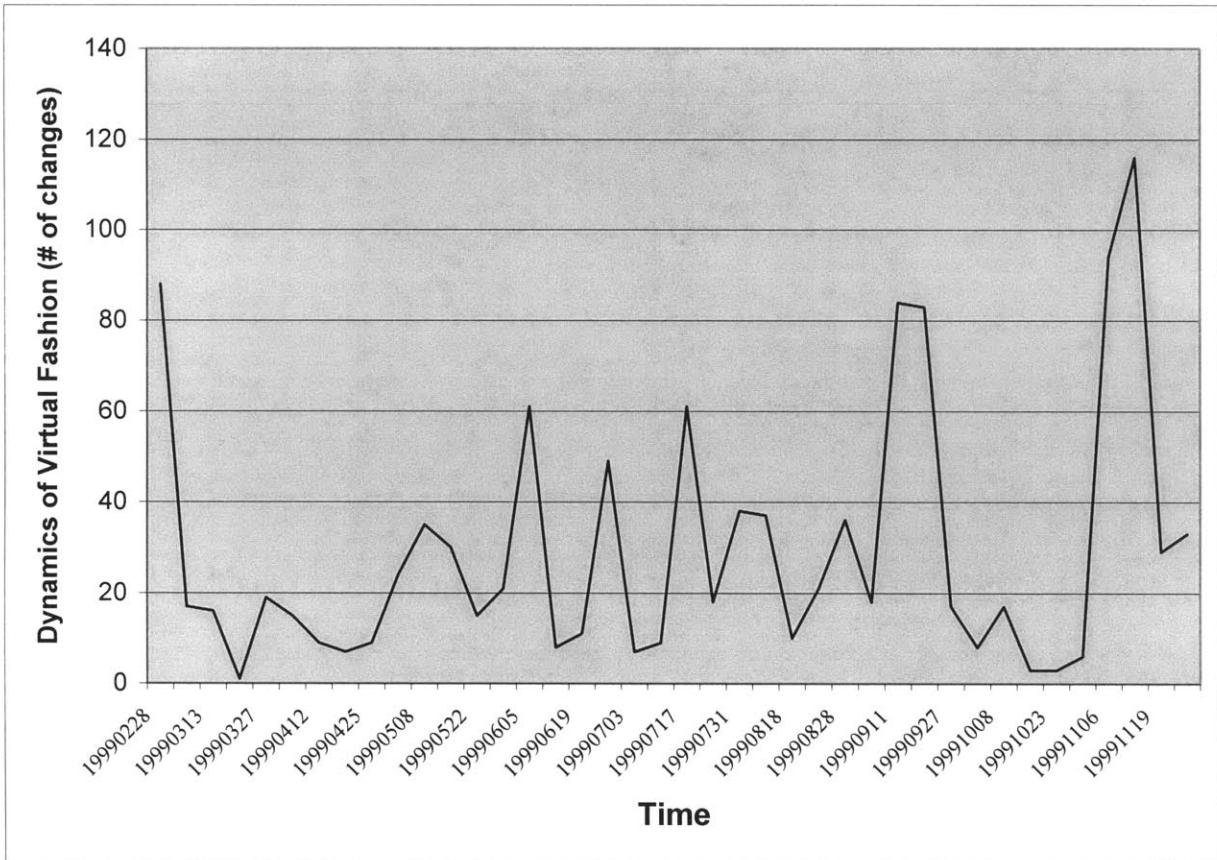


Figure 23: The dynamics of virtual fashion activity among Media Lab homesites.

Other statistics

The spatial-temporal archive can also be used for other analyses, which may provide valuable information about virtual communities. For example, below are some sample statistics of Media Lab homesites:

Total pages: 6804

Pages modified during the week of 2/14/1999-2/21/1999: 236 (3.5%)

Pages modified during the month of 1/21/1999-2/21/1999: 1834 (27.0%)

Pages modified time unknown: 224 (3.3%)

Total homesites: *159*

Homesites modified during the week of 2/14/1999-2/21/1999: *13 (8.2%)*

Homesites modified during the month of 1/21/1999-2/21/1999: *31 (19.5%)*

The above statistics suggest that in the Media Lab domain, about one third of pages are modified in that week, and about one fifth of homesites are modified in that month.

4.3 Case Studies

Innumerable interesting cases can be found using the system. The following are some instances:

Life and death of CHI99 in the Media Lab

Figure 24 shows the change of popularity of the CHI99 link

<http://www.acm.org/sigchi/chi99/> in the Media Lab. It became popular when people submitted papers to it and got accepted. After that, nobody cared about it anymore.

Finally, people began to unlink it.

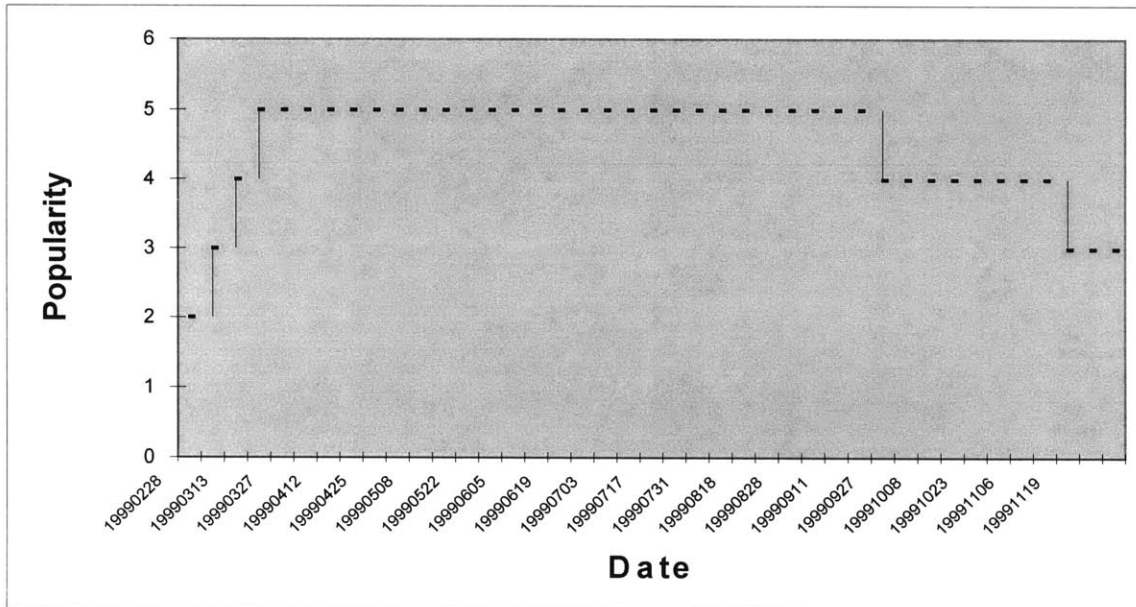


Figure 24: The popularity change of CHI99 in the Media Lab.

A Lively Group in the Media Lab

Figure 25 shows the change of popularity of the link to the Synthetic Characters Group <<http://characters.www.media.mit.edu/groups/characters>> in the Media Lab. The Synthetic Characters Group homepage attracted new members of the group and people who took the class offered by Professor Blumberg to link it on their homesites.

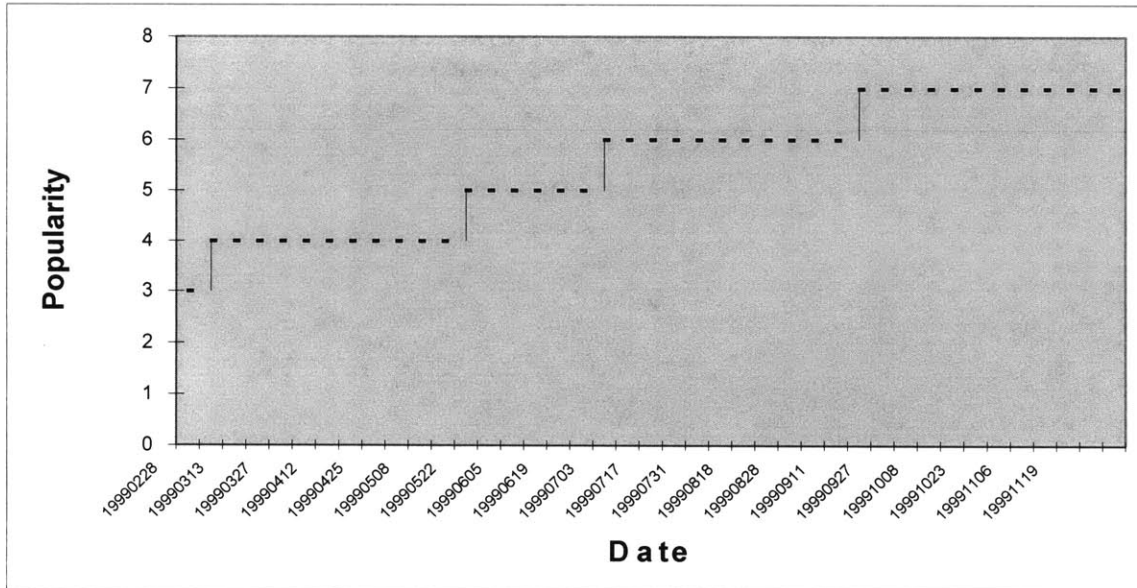


Figure 25: The popularity change of the Synthetic Characters homepage.

Who is Nikos Drakos, the most popular guy in the lab?

By 12/5/1999, 105 out of 159 Media Lab members linked to the Media Lab homepage, and 40 members linked to the MIT homesites. What is the third most popular link in the Media Lab? The answer is: Nikos Drakos' homepage, which is linked by 15 people.

Who is he? He is the creator of LaTeX2HTML, a popular tool used by many people to convert their theses from LaTeX format to HTML format. At the end of these documents, there is a disclaimer "About this document...", which points to Mr. Drakos' homepage. It is not getting more popular in the Media Lab in the past one year, possibly because fewer people are using LaTeX these days. This is an instance of how the tools people use to create their homepage may affect the observation of virtual fashion.

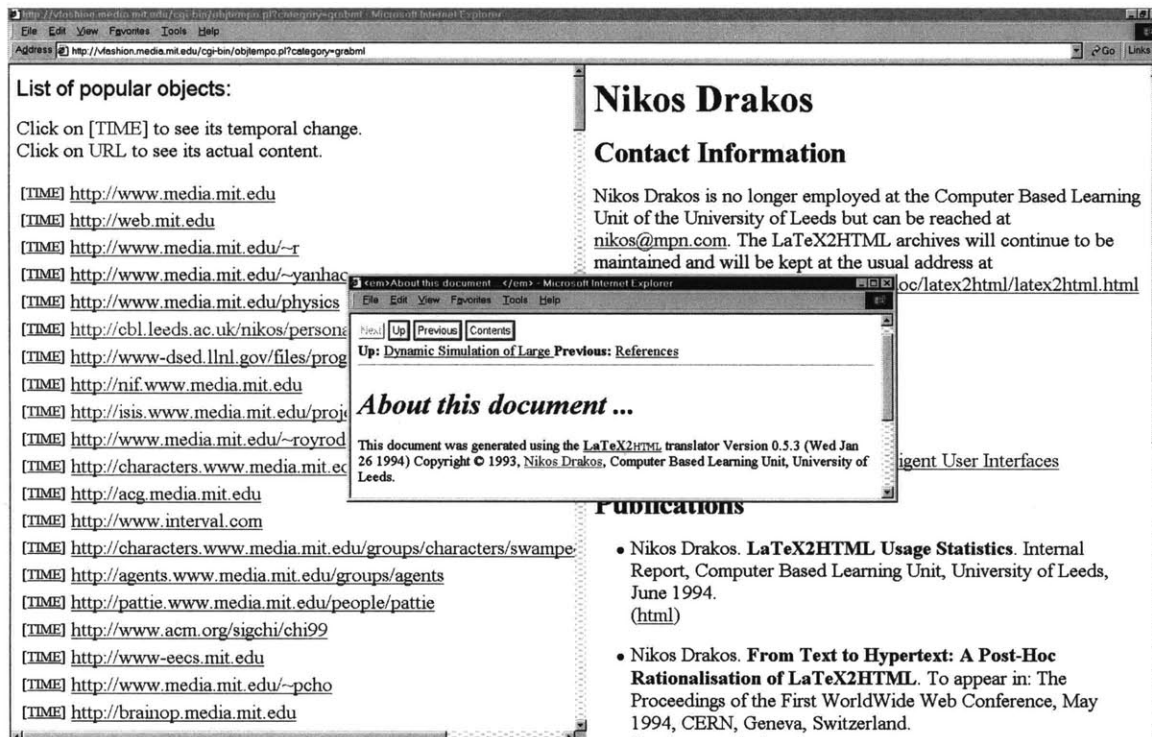


Figure 26: The creator of LaTeX2HTML plays an important role in the lab.

AddMe!, for a while

AddMe! <`http://www.addme.com`> is a free service that helps users submit their web site to thirty popular search engines and directories on the web. AddMe! asks users to put a link on their homesites in exchange for the service. As shown in Figures 27 and 28, people link and unlink it every now and then. Many people linked it in one week, and unlinked it in the following week.

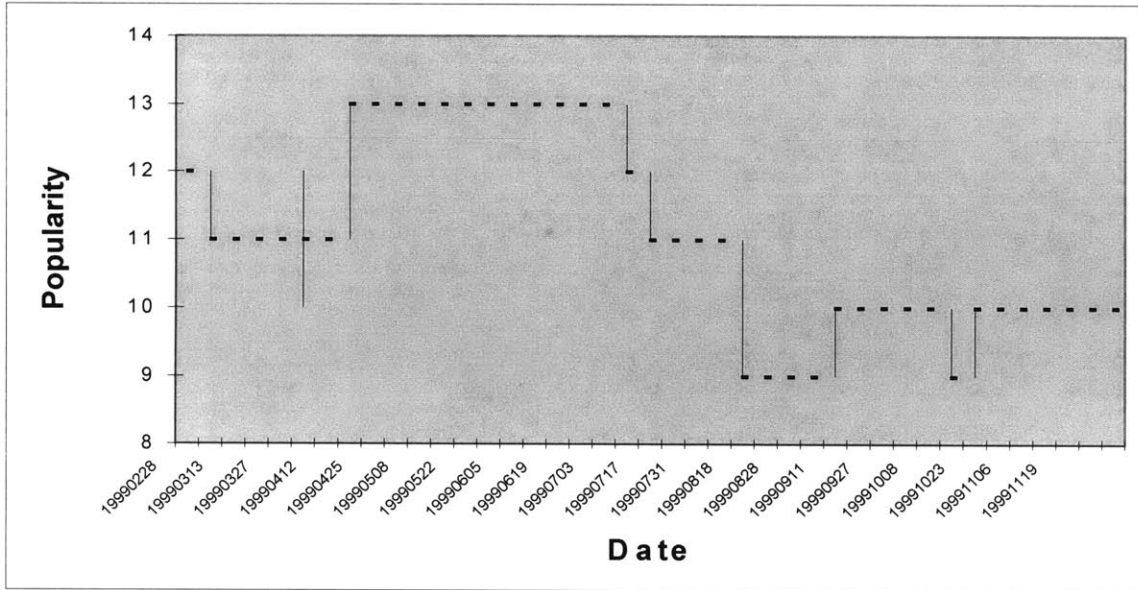


Figure 27: The popularity of AddMe! on Area51 oscillates.

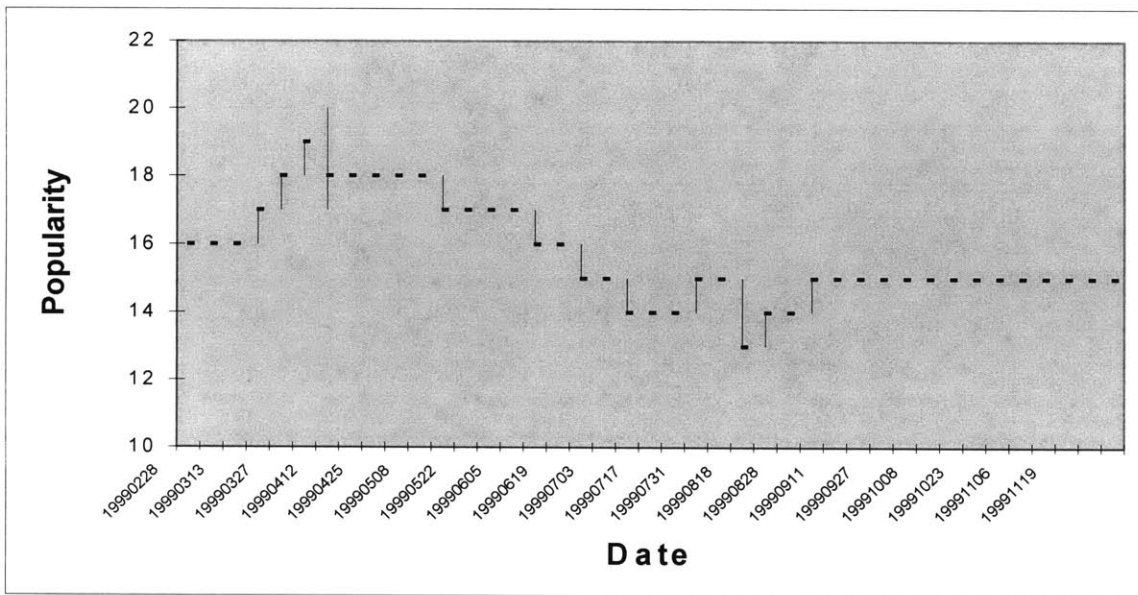


Figure 28: The popularity of AddMe! on Heartland oscillates.

StarWars never dies - it just fades away

"Star Wars: Episode I - The Phantom Menace" debuted in late May of 1999. On Area51, its web site <<http://www.starwars.com>> became more and more popular until early July,

and then started to decrease. Furthermore, nobody in the Heartland or MIT Media Lab ever linked to it during the same period.

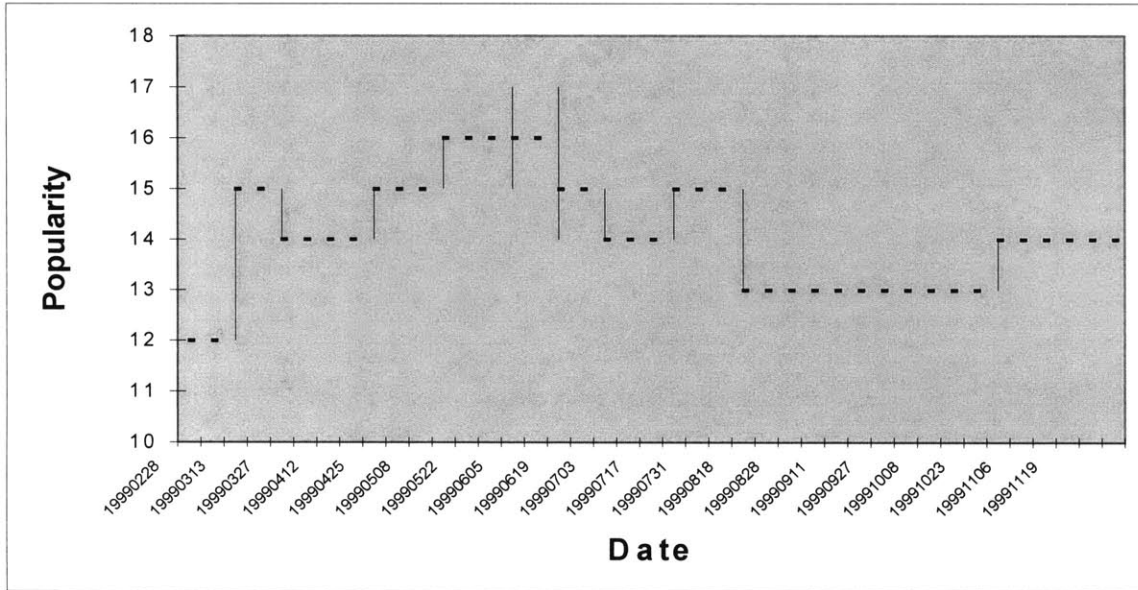


Figure 29: StarWars was a fad on Area51.

Do you Yahoo!?

Yahoo! is one of the most popular links almost everywhere. Yahoo! bought GeoCities in 1999 and then started to place links to Yahoo on GeoCities homesites since September. This is the reason that the number of links to Yahoo! increased dramatically on GeoCities since then, as shown in Figures 31 and 32. This is an example of how commercial operations affect the measure of virtual fashion.

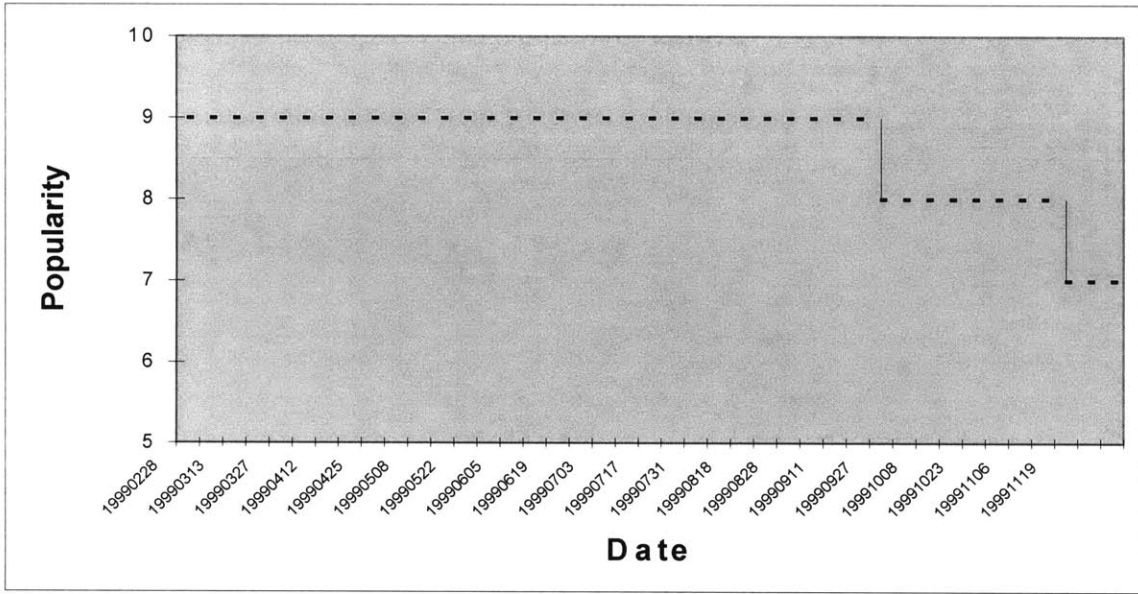


Figure 30: The popularity change of Yahoo in the Media Lab.

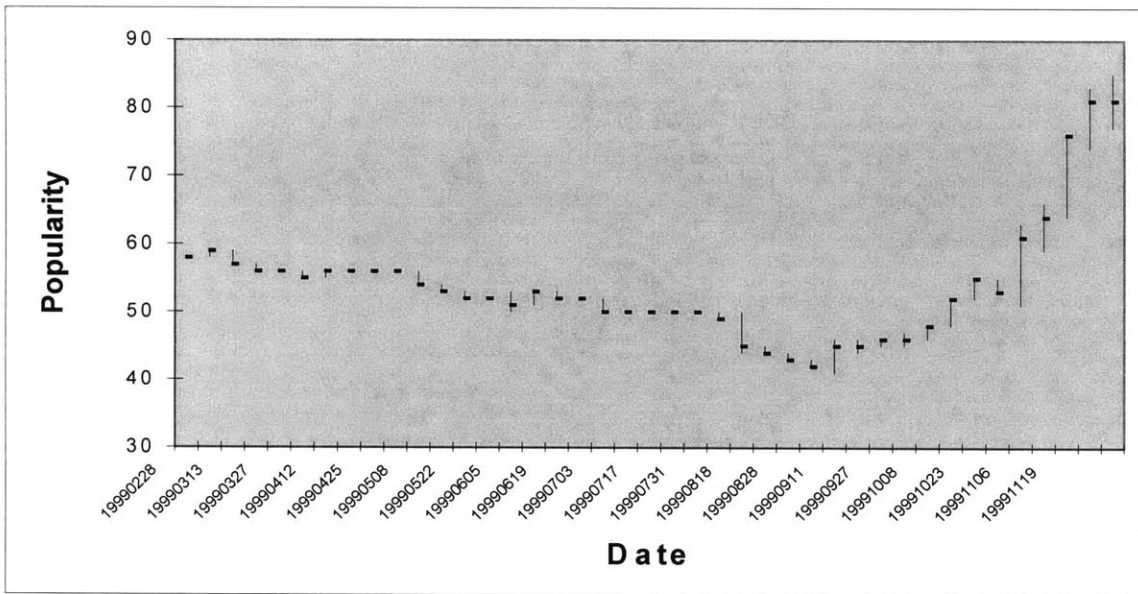


Figure 31: The popularity change of Yahoo on Area51.

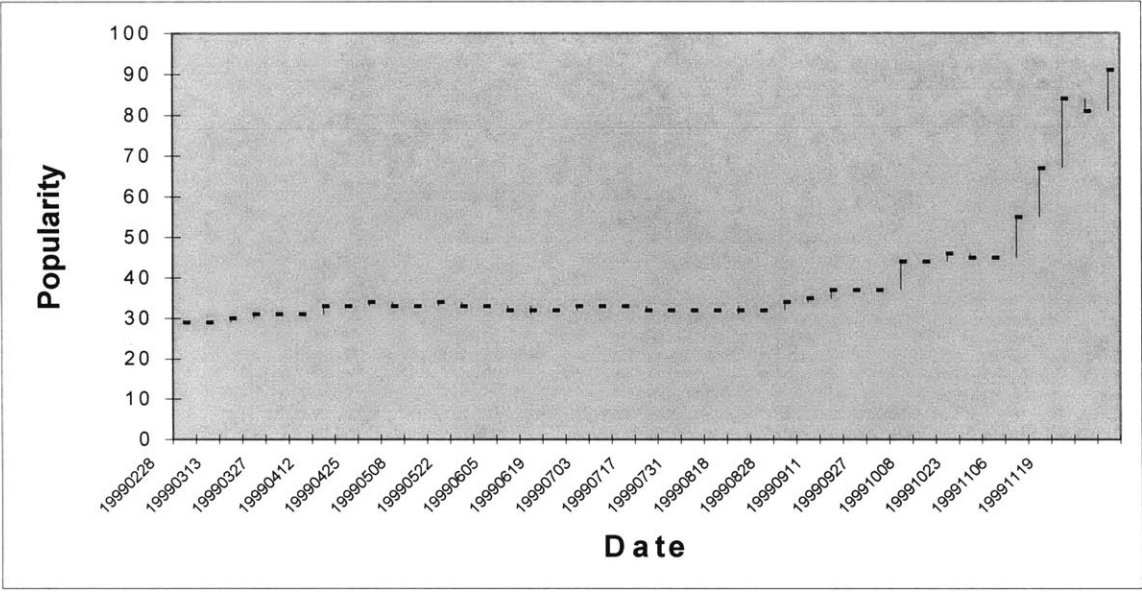


Figure 32: The popularity change of Yahoo on Heartland.

Revitalization of ICQ

As shown in Figures 33 and 34, the popularity of ICQ went all the way down till mid 1999, and then seemed to revitalize after that. It would be interesting to survey the reasons behind the revitalization.

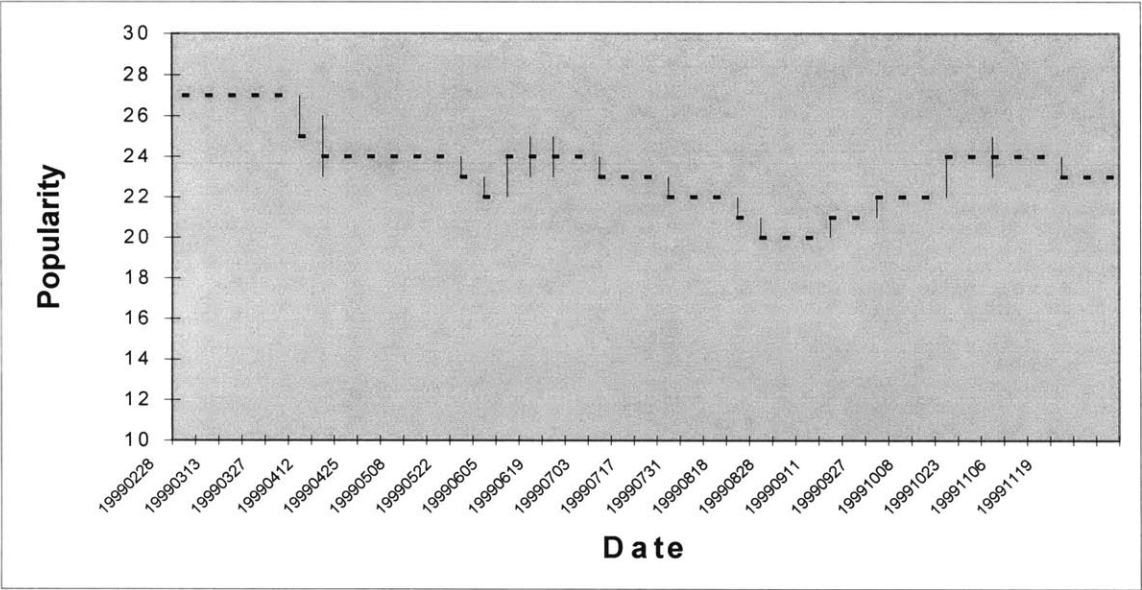


Figure 33: The popularity change of ICQ on Area51.

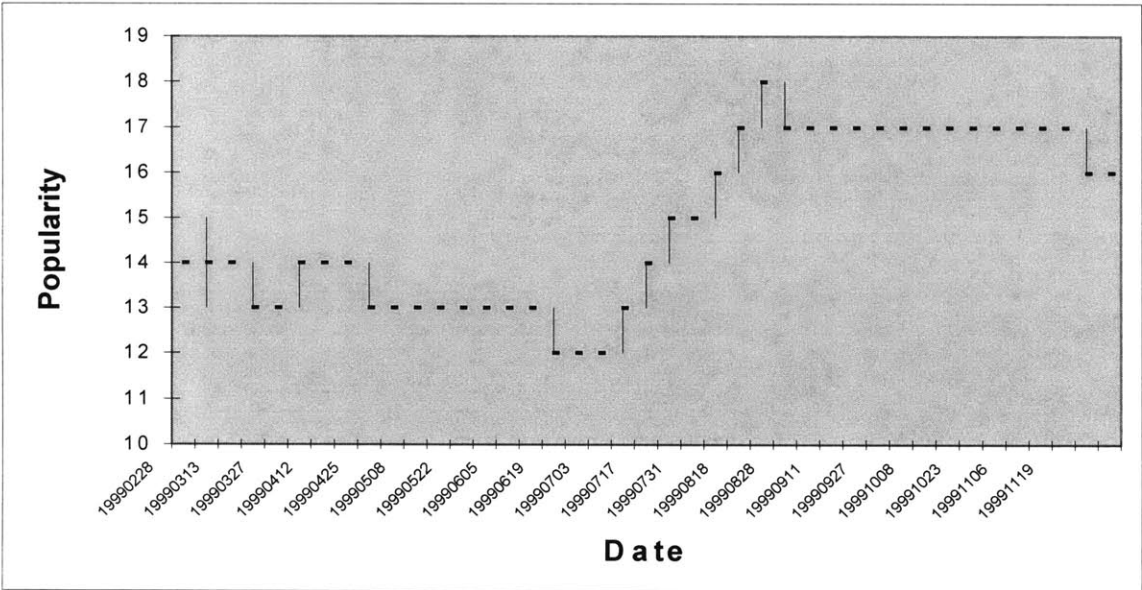


Figure 34: The popularity change of ICQ on Heartland.

4.4 Visualization

Visualizing virtual fashion is a difficult task when large numbers of homesites and virtual objects are involved. Instead of using existing systems such as the Self-Organizing Map (SOM, Kohonen 1997) or the approach of Visual Who (Donath 1995b), I have tried to design several visualization systems for virtual fashion. My first several designs using the simulated annealing approach failed to highlight the structure within virtual fashion or communities. Finally, I design the "Community Contours" visualization, which visualizes community structure based on the hierarchical clusters derived from the clustering subsystem. In the following section, I will briefly analyze the reasons the first several designs failed and describe the design of Community Contours.

Failed experience

I started out with a simulated annealing approach to visualize virtual fashion. Each object is assigned a "positive charge," and each homesite is assigned a "negative charge." Therefore, objects repel one another, and homesites repel one another, and objects and homesites attract one another. The amounts of charges on each object or homesite are determined by several factors. On the screen, objects are shown as red dots, and homesites are shown as white dots. In this way, I hope the objects and homesites will reach a steady state on their own and result in a Nebula-like image, which should be clusters of related data similar to the way galaxies form in the universe. This did not prove very useful as the data became extremely cluttered when a lot of homesites and objects get involved. Figure 35 shows two examples.

I tried to refine the design in various ways, including using different insertion algorithms to pre-arrange the objects and homesites before running the simulated annealing, changing the way charges are assigned, and shaking the result to make them anneal several times. All of these attempts failed to produce a satisfactory image that can reveal

multi-level group hierarchy. Figure 36 shows one of the best images, but the approach usually generates worse images, as shown in Figure 35.

Figure 37 is an image of SOM for comparison. It may not be more visually appealing, but at least it takes better advantage of screen space. Note that all the approaches, including SOM's and mine, will result in discontinuous images if we want to see the image of virtual fashion over time. This is a big problem if you want to see continuous images of virtual fashion over time. Some heuristics can partially solve the problem, but I have not found a perfect solution. A separate research will be needed to address this problem.

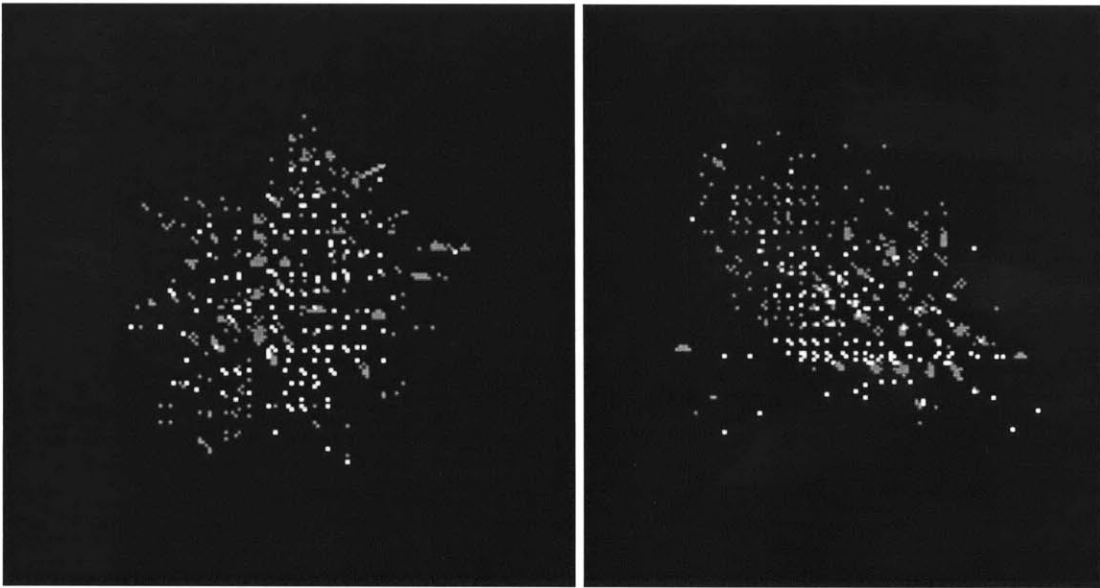


Figure 35: Two screenshots from the simulated annealing visualization system.

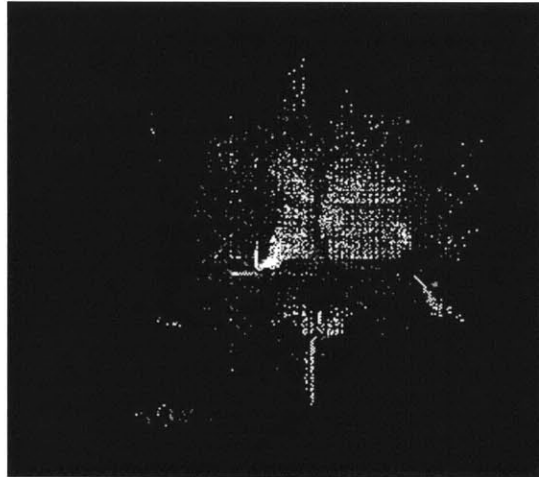


Figure 36: A refined image. But it does not generate this kind of images all the time. Most of the time, and images still look like those in Figure 35.

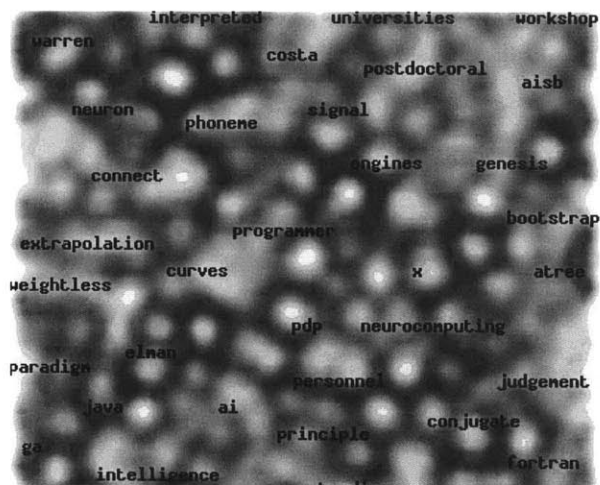


Figure 37: A screenshot of the Self-Organizing Map.

Thus, I changed the design to the "Community Contours" visualization, which differs from previous trials in two major aspects. First, to make the algorithm less complicated, objects and homesites are no longer mixed together on the screen. Instead, there is one picture for object hierarchy, and another picture for homesite hierarchy. Second, all the objects or homesites are clustered before being visualized, instead of using simulated annealing to let them cluster on their own. This was the reason that I developed the clustering subsystem. In other words, "Community Contours" get its input from the clustering subsystem.

Community contours

"Community Contours" is a design to visualize communities of people or virtual objects. In this design, circles represent clusters, and circles of parent clusters surround circles representing child clusters. Basic heuristics for positioning the circles are also discussed in the section. This design helps to visualize hierarchical community structures in an appealing way.

"Hierarchical clusters" in general means a hierarchical structure of groups of objects. Roughly speaking, the objects are clustered recursively according to the similarity between them (Everitt 1980). Traditionally, hierarchical clusters are shown with tree representation as the example in Figure 38 shows.

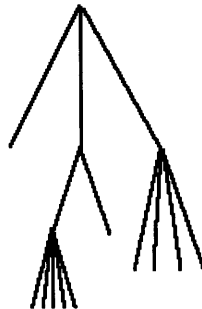


Figure 38: An example of tree representation of hierarchical clusters.

With the emergence of virtual communities, there are more and more applications to visualize hierarchical structures of a community. In contrast to real-world communities, virtual communities are hidden under layers of unintuitive text and need to be visualized so that they can be made more apparent. Therefore, representing the structure in ways other than the traditional tree representation, which may not be visually appealing for end users, is of importance.

To this end, I introduce the design of Community Contours to enable users observe and explore the underlying social structure within a community.

Next I will present the basic design of the visualization, describes the way it positions clusters, and discuss the trade-offs of this approach. Examples of applications are also described.

Representation

The basic unit of this visualization is a circle. A circle represents a cluster; the circle of a parent cluster surrounds all circles of its child clusters. The size of a circle, by default, is proportional to the number of its members. Brightness within the area of a circle represents the relational tightness within the cluster; tighter related clusters are of brighter color. Figure 39 is an example.

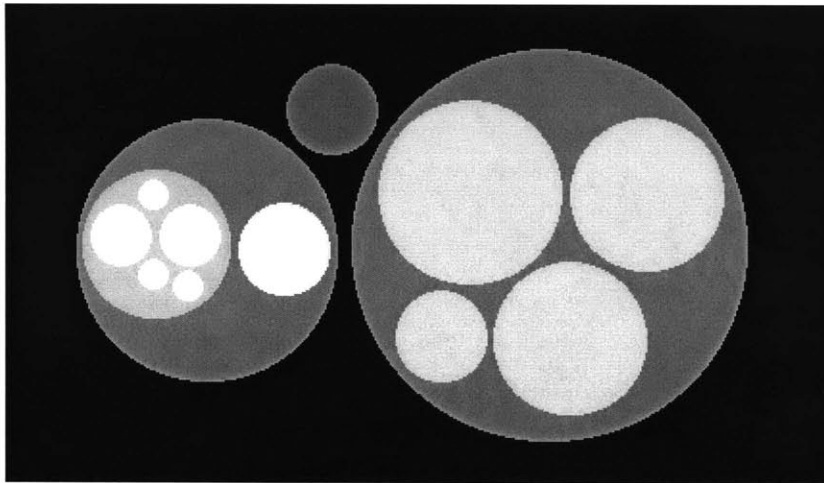


Figure 39: A basic layout of Community Contours.

Positioning

It is a challenge to develop heuristics for automatically positioning these circles. My approach is to assume the size of the screen is infinite at first, and then start by inserting the level farthest from the root. That is, circles of child clusters are inserted earlier than circles of their parental clusters. This is accomplished graphically by surrounding child circles by parent circles until the root of the hierarchy is reached.

On each level, circles of clusters are positioned one after another with larger circles placed earlier. First, the largest circle is placed in the center of the screen or next to previously inserted circles of the former level. Then, the second largest circle is placed next to the largest one. The third largest circle will be inserted next to the largest or the second largest circle in a position that minimizes the required size of the parent circle that encircles all these circles. All subsequent circles will be placed in the same manner that minimizes the area occupied by the circles. Finally, a parental circle encircling all these circles is drawn. After all levels of clusters are positioned, the image will be zoomed to fit the actual size of the screen.

This inside-out approach of placing circles avoids difficult algorithm problems of Circle Packing (Wolfram 2000, Stephenson 2000, Friedman 2000). If placing circles outside in, the program would have to calculate all the radius of the inner circles algorithmically, which is a NP-complete problem. Also, if you want the computer to pack many circles aesthetically, a lot of mathematical studies need to be done in advance. Given the limited time, what I can do is to propose the design and have a simple version running instead of developing an aesthetic and well-balanced one, which may be another NP-complete problem.

The member objects of each cluster will be placed inside the circles to enable users perform operations on them, such as clicking on an object. The actual representation of a member object can be a grid, a dot, an icon, etc.

Threshold

If there are too many groups and thus contours, the image may become too complex. A solution is to implement a threshold for the relational aspects of the data. In this way, when there are too many groups being differentiated due to their relational differences, a user can simply increase a threshold setting to merge less related groups so as to decrease the number of groups cluttering the screen at one time. This method is very useful when the visualization is applied to large amounts of data. To automatically set the default

threshold that may result in a visually appealing image, an approximate function can be developed.

Dynamics

Live communities are always changing. Such changes may be illustrated by animating a series of hierarchical clusters over time. Each new data set collected is visualized and added as a slide to the temporal animation. This allows users to see changes of the social structure by observing the way contours and objects move and transform in the animation.

However, since the result of hierarchical clustering analysis sometimes varies a lot with only a little difference and thus makes the picture changes too reactive, further research is needed to smooth adjacent visualization by having the clustering algorithm take into account previous data. This is an inheritance problem no matter what design is used to visualize temporal data such as virtual fashion. I will discuss this in more detail later.

Trade-offs

Using circles is not the most space-efficient way, but looks better than using rectangles or traditional tree-like structures. Squares are not as visually appealing or intuitive as a less rigid structure similar to contour maps. A set of squares within squares within squares is hard to parse and one can easily see illusory rectangles formed in the negative space of the layout, but the circles are easier to parse visually since the positive and negative space have different contours. So far, the circles have proved to be successful at showing distinctive hierarchy in the group and sub-group classes.

An issue that some people may suggest with this visualization technique is the possible illusion caused by certain objects that visually appears to be close on the screen, but are in different parent clusters so are not necessarily more related to each other. However, unless I adopt an approach such as SOM, which emphasizes neighboring relations rather than cluster hierarchy, I have not found any universal solution that perfectly defines closer

objects to mean more related objects. Methods like SOM are helpful for visualizing neighboring relations but are not designed for visualizing hierarchical clusters.

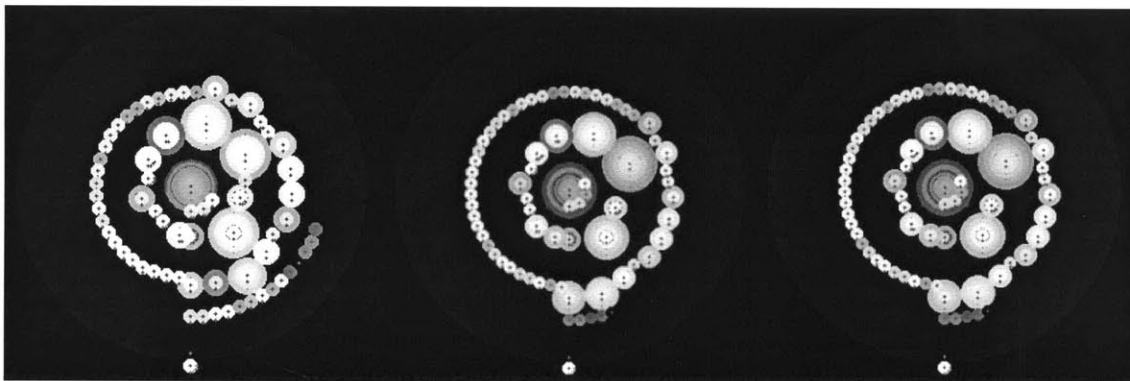
Applications

For the virtual fashion project, the strategy of the design is three-fold: On a large-scale, the visualization shows the general community structure of the whole set of homesites using map contours for intuition. On the medium scale, it reveals how a homesite fits into the ensemble. The small-scale visualization allows users to easily traverse the homesites or objects in them in the context of cultural dispersion and other properties.

Researchers may also use this visualization to represent the community structure of consumers or even products. Instead of using a tree to represent the data, Community Contours provide a novel approach that may look more interesting and allow users to obtain information that may not have been as intuitive in a rigid tree structure.

Snapshots

Figure 40 is a series of snapshots of the object hierarchy in the MIT Media Lab. The snapshots are placed from the upper-left (earliest) to the bottom-right (latest). Some snapshots are obviously not continuous from that of its previous week. All of these cases are the outcome of the drawback of the hierarchical clustering algorithm: a slight change sometimes results in significant changes of the hierarchy.



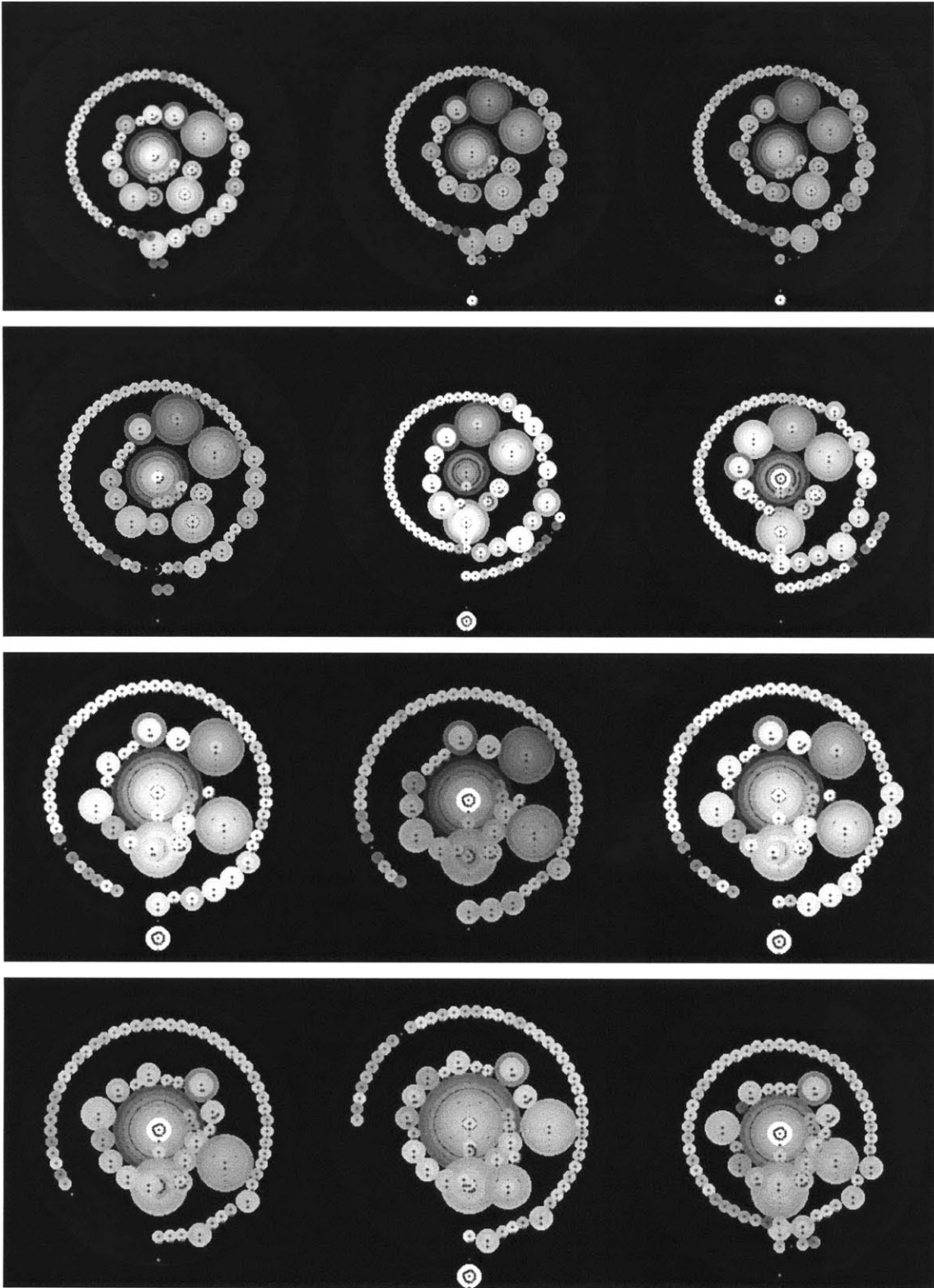


Figure 40: A series of snapshots for the object hierarchy in the MIT Media Lab.

Ideal and Reality

Ideally, visualizing virtual fashion over time may reveal something about how the role individuals play changes with the change of virtual fashion. An analogy in the real world is that if twenty years later, you wear the same style of clothing as you wear today, even if the style is in fashion right now, you may look bizarre at that time. Thus, even if somebody does not change at all over time, his role regarding fashion diffusion will be changed a lot. An ideal visualization should highlight this phenomenon. The list of this kind of homesites can be easily found by an algorithmic approach, but to highlight them in an intuitive visualization is not easy.

The difficulty of the visualization lies in not only the design of the visualization, but also the heuristics to smoothly animate the visualization over time. Out of intuition, people may think that they can just put together the visualization of different weeks and then get the animation. But after lots of trial-and-error, I found this approach failed to generate smooth animation. Because there are more and more virtual objects over time and the relatedness between homesites or objects changes every week, most algorithms cannot generate pictures that differ only slightly when only a little change occurs in the data set. Even a few new objects' network effect makes the result of hierarchical clustering; Community Contours, and SOM change a lot. This does not necessarily ruin the animation, but makes it far from satisfactory. Especially when many homesites and objects are involved in the visualization, the discontinuity makes it more difficult to observe how a certain home site's role changes in the community.

This discontinuity is not an unsolvable problem, but will take much more time to work out. Conceptually, we may feed data of all the weeks into the system and treat them as a whole instead of treating the data of different weeks separately. This can be a future area of research.

Another way to avoid the problem is to display only a slice of data at the same time, instead of generating a global picture for all the data. VisualWho (Donath 1995b) is based on this approach.

5. Discussion on Theoretical Modeling

Although there are several existing Web-based data-mining services based on hit-rate, link information, or textual context, none of them shows the temporal dynamics of the popularity of virtual objects, nor do they reveal the role individuals play in the diffusion of online culture. In contrast, my system can help find out which homesite adopts which popular object at what time and for how long. This information lays the groundwork of further research in theoretical modeling of people's interaction pattern on homesites.

Potentially various types of modeling can be done on the collected data. For my master's research, I have tried to touch two fields: high-level modeling of virtual fashion, and low-level modeling of cultural transmission among homesites. This work is not vigorous and complete sociological research but can provide important insight into the virtual world.

5.1 Background Research

For centuries, sociologists have been observing fashion change and thinking of models behind it. But it is difficult to verify them quantitatively in the real world (Davis 1992).

The key idea of the theoretical modeling is that fashion is the temporal change in the social meaning of an object. In a social hierarchy, the forces of differentiation and emulation, according to the “trickle-down theory”, possibly fuel these changes. There can be many such hierarchies in a culture. In a highly mobile culture in which information flow is rapid, changes in fashion will also be rapid. Furthermore, the flow of fashions reveals a lot about the social structure of a society: it shows where there is contact between individuals and delineates the sub-cultural hierarchies. These sub-cultures are

developed by people who have similar runs of experience through common interactions, according to the "collective selection theory."

5.2 Modeling Framework

With the collected data of virtual fashion, it is feasible to verify some sociological theories in the virtual world. It usually involves several steps. The first is to find online analog of a theory. The second is to map the theory into a model that can be verified. Then, data collected by my system can be used to verify the model. Although the mapping process may make the model a little different from the original theory, convincing results can still be achieved with a carefully designed process.

For instance, I have verified that the original trickle-down theory without modification does not hold online, but the revised versions of trickle-down theory may still be helpful. Among the eight thousand homesites I chose to track, I could not find a unified hierarchy. This result corresponds to my prior analysis on the pluralism and polycentrism of fashion online, as described in the "Sociological Issues" chapter.

5.3 Verification of the Trickle-Down Theory

The original trickle-down theory is based on the hypothesis that there is a universal hierarchy of social status in the society, and people with lower status usually follow fashions that have already been adopted or created by people with higher status. If this hypothesis is not true among homesites, then the trickle-down theory in its original form cannot be used to describe people's patterns of interaction on the Web.

Strictly speaking, there are various kinds of fashion on the Web, and link fashion is one kind of them. Furthermore, part of the link fashion may just be coincidence or may result from other channels of mass communication but not people's interaction online, and the other part of link fashion is the outcome of people's interaction. If we can verify that there

is no "universal hierarchy" even among only eight thousand sampling sites, we know that the original trickle-down theory cannot describe the mechanism behind link fashion.

I have written programs to analyze the data and found that there is definitely no universal hierarchy regarding the diffusion of fashion among the eight thousand homesites from February 1999 to February 2000. That is, the trickle-down theory implies that if A adopts a fashion earlier than B, and B adopts the fashion earlier than C, then for another fashion, the diffusion of fashion will likely be A->B->C, or A->C. This is not the case online. Instead, C->B->A or C->A may happen, too. This may also result from factors like fashion pluralism and polycentrism, as described in the "Sociological Issues" chapter.

Now we know there is no universal hierarchy. Are there individual-to-individual diffusion chains? In other words, are there guys A and B; B usually follows fashions later than A? My analysis of fashion diffusion in the MIT Media Lab shows that there may be some such cases, but they are not significant enough to be a general phenomenon. For example, a student follows objects his advisor linked to for couples of times, but most other people in the Media Lab do not do so. A through study on a more complete data set will be needed study if those revised versions of trickle-down theory can model fashion online.

Figure 41 shows the figures behind the analysis of individual-to-individual diffusion in the Media Lab domain. The format of each line is: "A, B: {O | O are objects that B adopts after A}." If {O} contains more than one element, it is possible that B tends to adopt objects later than A. It is also possible that B linked to those objects later than A by coincidence. The more elements of O, the more chance that the diffusion is not by accident. i.e. the "chain" between A,B is more certain.

To justify if the diffusion is by accident or not, a helpful and accurate way is to manually inspect those objects and homesites, as what I have done on this data set. My inspection shows that most of the cases are just by accident, and some of them are likely to be

consequences of interaction between individuals. Note that the latter provides some basic clues for predictive modeling.

But the Media Lab data set is not representative enough of cultural transmission online because there is not a great deal of cultural diffusion among these homesites, as mentioned in the "Output and Case Studies" section. To have a more complete result, analysis of a larger data set, or at least of the Area51 and Heartland data, is required.

Manual inspection takes a lot of time although it is accurate. To analyze the data set of Area51 and Heartland, a well-designed computational approach with statistical hypotheses is needed. Even though a complete analysis is beyond the scope of this thesis, I have roughly concluded some of the statistical conditions as follows:

In the example of

$$F::I\{I \mid \text{initial set}\}::A1,A2::B::C1,C2,C3::D::B$$

where

- F is the URL of the popular object.
- I is the set of homesites that had linked to F before the system started tracking. (Therefore, we don't know the sequence of adopting F among {I}.)
- A1, A2 are homesites that linked to F during the first week the system started tracking.
- B is a homesite linked to F during the second week.
- C1,C2,C3 are homesites linked to F during the third week.
- D is a homesite linked to F during the fourth week.
- B linked to F again during the fifth week. That is, there is one more link on homesite B that linked to F during the fifth week.

Regarding the diffusion chain between all the involved individual, it is possible that

1. {I} are of higher status than {A1,A2,B,C1,C2,C3,D}.
2. S is of higher status than {A1,A2,B,C1,C2,C3,D}.

3. A is of higher status than B, with interval equal to 1 week; B is of higher status of D, with interval equal to 2 weeks. Etc...
 4. A1 can be of higher, equal, or lower status than A2, with interval less than one week.
 5. C1 is more likely a follower of A than D. That is, the longer the interval, the lower the certainty of the diffusion chain.
 6. The higher the total number of homesites that have already linked to F, the lower the certainty of the diffusion chain.
 7. Even though B linked to F during the fifth week, B is probably not of lower status than {C1,C2,C3,D} because B already have a link to F during the second week.
- and so on.

With all these conditions, it is possible to find out the certainty of the "chains" between individuals. In this way, human inspection will be needed only to verify those chains with high certainty. I hope that some day researchers can take advantage of these conditions as well as the data to analyze a large data set.

To sum up, my observation suggests that the hierarchy of homesites regarding the diffusion of virtual fashion is as follows:

1. As being the basic hypothesis the clustering subsystem, there are different groups of homesites. They form different sub-cultures of interest. Members of each group may also change their interests over time and thus leave the group.
2. The uniform and multi-layered class hierarchy described by the original trickle-down theory is not the case online. Other than some well-known virtual fashion initiators, the "trickle-down" effect is weak and the social structure is fragile between homesites. A thorough study on a large data set will be required to show if the modified theories can hold online.
3. There are some individual-to-individual chains, but I have not found this to be a significant phenomenon.

6. Conclusion

I have examined and verified that the cultural phenomenon of virtual fashion, with the related issues of sub-cultural delineation and so on, does exist online, at least in the area of about eight thousand homesites that I have explored. This can be validated by using the online service I have developed. In the process of verifying virtual fashion, I built a system and developed the algorithms for measuring virtual fashion on the net. The system and methods I have developed are in themselves important and practical tools.

This interdisciplinary research touches areas including system design, algorithm design, data mining, and sociology. There are several contributions of the research. First is the development of the temporally consistent web-page retrieval method, the efficient object/homesite indexing method, and so on, to automatically track and analyze virtual fashion. Second is the analysis and examination of significant web phenomena, such as the different ecology in different virtual communities. Finally, the theoretical modeling opens a new door to the research of online culture and provides deeper understanding of the culture of the Web.

7. BIBLIOGRAPHY

ADscience 1999

ADscience Ltd. ADfilter, 1999.

Available At <http://www.adfilter.com/>

Bell 1947

Bell, Quentin. On Human Finery. London: Hogarth Press, 1947.

Berners-Lee et al. 1996

Berners-Lee, T., Fielding R., and H. Frystyk. "Hypertext Transfer Protocol -- HTTP/1.0." RFC1945. May 1996.

Available At <http://www.w3.org/Protocols/rfc1945/rfc1945>

Blumber 1969

Blumer, Herbert. "Fashion: From Class Differentiation to Collective Selection." Sociological Quarterly, 10(3), 1969.

Botafogo et al. 1992

Botafogo, R., Rivlin, E., and Shneiderman, B. "Structural analysis of hypertext: Identifying hierarchies and useful metrics." ACM Trans. Information System. Vol. 10, 1992.

Brightwater 1999

Brightwater Software. Drag-and-Filter, 1999.

Available at <http://www.brightsoft.com/products/dnf/>

Chiou and Donath 2000

Chiou, Ta-gang and Donath, Judith. Inferring Sub-culture Hierarchies on the World Wide Web. Proc. IEEE ICDCS International Workshop of Knowledge Discovery and Data Mining in the World-Wide Web, Taipei, Taiwan, April 2000.

Csikszentimihalyi and Eugene 1981

Csikszentimihalyi, Mihalyi and Rochberg-Halton, Eugene. The Meaning of Things: Domestic Symbols and the Self. Chicago: University of Chicago Press. 1981.

Davis 1979

- Davis, Fred. *Yearning for Yesterday, a Sociology of Nostalgia*. New York: Free Press, 1979.
- Davis 1992
Davis, Fred. *Fashion, Culture and Identity*. Chicago: University of Chicago Press. 1992.
- Donath 1995a
Donath, Judith. "Sociable Information Spaces," presented at the Second IEEE International Workshop on Community Networking, Princeton, NJ, June, 1995.
- Donath 1995b
Donath, Judith. "Visual Who." Proc. of ACM Multimedia '95, Nov 5-9, San Francisco, CA.
- Everitt 1980
Everitt, B. "Cluster Analysis." Halsted, NY, 1980.
- Fielding et. at. 1999
Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. "Hypertext Transfer Protocol -- HTTP/1.1." RFC 2616. June, 1999.
Available At <ftp://ftp.isi.edu/in-notes/rfc2616.txt>
- Friedman 2000
Friedman, Erich. "Circles in Circles."
Available at <http://www.stetson.edu/~efriedma/cirincir/>
- Gibson et al. 1998
Gibson, D., Kleinberg, J., Raghavan, P. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.
- Horowitz 1975
Horowitz, R. Tamar. "From Elite Fashion to Mass Fashion." Archives Europeennes de Sociologie, I6(2).
- Kessler 1963
Kessler, M.M. "Bibliographic coupling between scientific papers," American Documentation, Vol. 14, 1963.
- Klapp 1969
Klapp, Orrin. *Collective Search for Identity*. New York: Holt, Rinehart and Winston, 1969.
- King 1963

- King, Charles W. "Fashion Adoption: A Rebuttal to the 'Trickle-Down' Theory." Toward Scientific Marketing, ed. Stephen A. Greyser, Chicago: American Marketing Association.
- Kohonen 1997
Kohonen, T. "Exploration of very large databases by self-organizing maps." Proceedings of ICNN'97, International Conference on Neural Networks, 1997.
- McCracken 1988
McCracken, Grant. 1988. Culture and Consumption: New Approaches to the Symbolic Character of Consumer Goods and Activities. Bloomington: Indiana University Press, 1988.
- O'Hara 1986
O'Hara, Georgina. The Encyclopedia of Fashion. New York: Abrams. 1986.
- Pirolli et al. 1996
Pirolli, P., Pitkow, J., Rao, R. "Silk from a sow's ear: Extracting usable structures from the Web." Proc. ACM SIGCHI Conference on Human Factors in Computing, USA. 1996.
- Pirolli and Pitkow 1997
Pirolli, P. and Pitkow, J. "Life, Death, and Lawfulness on the Electronic Frontier." Proc. 1997 Conference on Human Factors in Computing Systems, ACM, Los Angeles, CA, USA. 1997.
- Polegato and Wall 1980
Polegato, Rosemary and Wall, Marjorie. "Information Seeking by Fashion Opinion Leaders and Followers," Home Economics Research Journal. Vol. 8, May, 1980.
- Rivest 1992
Rivest, R. "The MD5 Message-Digest Algorithm." Internet Activities Board, RFC 1321, April 1992.
- Rogers 1983
Rogers, Everett M. Diffusion of Innovations, third ed. New York: The Free Press, 1983.
- Siemens 1999
Siemens Computer System Development. WebWasher software, 1999.
Available at <http://www.webwasher.com>
- Simmel 1904
Simmel, Georg. "Fashion." Reprint In American Journal of Sociology 62 (May 1957), 1904.

Small 1973

Small, H., "Co-citation in the scientific literature: A new measure of the relationship between two documents." J. American Soc. Information Science, Vol. 24, 1973.

Stephenson 2000

Stephenson, Ken. "Circle Packing."

Available at <http://www.math.utk.edu/~kens/>

Terveen and Hill 1998

Terveen, L. and Hill, W, "Finding and Visualizing Inter-site Clan Graphs." Proc. 1998 Conference on Human Factors in Computing Systems, CHI 98. ACM, New York, NY, USA, 1998.

Wolfram 2000

Wolfram Research. "Circle Packing."

Available at <http://mathworld.wolfram.com/CirclePacking.html>