論文の内容の要旨

Title A Study on Web Mining Techniques for Off-Line Enhancements of Web Sites
（ウェブサイトオフライン改善のためのウェブマイニング技術に関する研究）

Rios Sebastian A.
リオス セバスチァン アレハンドロ

WWW(World Wide Web)は20世紀において実現された最も大きな成果のひとつであるといえよう。今日、Webなしにコンテンツビジネスやエンターテイメントサービスを行うことは不可能な状況となっており、近年、企業でも考え方に変化が見られ、単にオンライン広告としてWebサイトを開設したり、ある種のファッションとしてWebサイトを開設するといった態度から、ビジネスを遂行するための新しく、大規模で簡潔な手法としてWebを捉えるようになっている。

毎日、たくさんの人々がインターネットに接続し、その中にはblogやVlogそして自身のWebサイト等を使ってWWWを上手に利用している人も少なくない。彼らのうちの恐らく全員が検索エンジンを用い、また、その中にはWWWを通じて製品やサービスを購入したり、自宅のリビングから映画を借りたり、ビデオ会議クライアント（OpenWengo, Ekiga等）を使ってインターネットを通じてビデオ会議を行ったりしている人もいる。あるいは、家族や友人と非常に低価格もしくは無料でコミュニケーションを取っている人もいる。このような背景からインターネット上のサービスの数はインターネット利用者数と同様に爆発的に増加している。

WWWはインターネットとHTML(Hyper Text Markup Language)の結合によって創出した技術であるが、この非常にシンプルなアイディアにより、誰もが自身の文書を発信できるようになったのである。一昔前は単純なHTMLでのプログラミングを行い、そのファイルをサーバにアップロードしていたが、その後、特定のツールを使うことにより、HTMLプログラミングやFTP(File Transfer Protocol)その他のことを何も知らなくても、非常に簡単にWeb文書を作成できるようになった。これによって、Webページ、Webサイト、Webポータル等の大増加が生じ現在に至っている。しかしながら、これは同時に、ラベル化も構造化もされていない、いわば役立つ情報を探し出すのが極めて難しい情報が大量に生み出されている、という事態を招いている。さらに、たとえユーザが興味深いWebサイトにたどり着くとしても、それを閲覧する際にどうしたら迷子にならないのかを知る手段がないという状況になってしまっている。

今日、WWWはいわば「文書の山」であり、何らかの情報を探そうとする人誰もがまず最初に莫大な無益な情報の山から有益な情報を探すことから始めなければならない。この状況はインターネットの普及がこのまま続けば年々ますます悪くなってゆくことは容易に予想される。一方、これらの問題を解決するためにセマンティックWebの開発も進められているが、今後10年でそれが立ち上がることは期待できないのが実情である。この新しいWebシステムはユーザに対しより簡単な方法で大量の情報にアクセスする手段を提供し良い検索結果を得られるようにするものであり、各Web文書本体のみならず、そのデスクリプション（メタデータ）をもXML形式で記述しようというものである。このため、セマンティック検索エンジンのようなソフトウェアでこの情報を読み取り、検索者の希望する検索結果をより関連の深い検索結果を返すことが可能となるが、上述の通り本技術は今日明日に使えるようになるものではない。

システム管理者ならびにWebを開設する組織にとって最大の課題は旧来からの顧客を維持し、新しい顧客を獲得するために “優れた” Webサイトをどのように立ち上げるのか、あるいは別の言い方をすると、継続的に顧客がこのWebサイトに戻ってくるようにするために顧客に価値ある情報を与えられるサイトをどのように立ち上げるのかということである。これはビジネスの視点に立った場合最も基本的な問題のひとつであるといえる。と言うのも、新しい顧客を得るコストは既存の顧客を維持するよりもずっとコストを要するからである。

この課題は解決が極めて難しいものであるが、この課題は以下の問題に分割することができる。一つ目は、ユーザが混乱することなしに簡単にサイト内の情報にアクセスできるビジ

ュアルデザインをどのように行うかということである。そして2つ目はユーザが必要な情報、製品、サービスをどのように与えるのか、ということである。そして３つ目に簡単かつ不明瞭でないブラウジングのためにWebコンテンツの構造をどのように改善するのか、ということである。

Nielsen氏はWebサイトにおけるユーザビリティの問題について論じており、また、Webサイトの有効性や効用がユーザビリティと強く相関している、という論を展開している。このとき、Webインターフェースに加え、サイトの構造やコンテンツを改善することによりユーザのブラウジングを効率的にすることも可能である。ユーザは自身である範囲のWebページ内で（あるいはある回数内のクリックで）必要な情報を見つけられない時、別のサイトに飛んでしまうのが常だからである。

これまでWebのサイト構造ならびにWebコンテンツを改善することによりWebサイトのユーザビリティを向上させることができるという研究事例はいくつか存在している。しかしながらこれらの問題は解決方策の良し悪しが通常主観評価で行われているという点である。また、別の研究事例として、Webサイトのコンテンツや構造自体をWebマイニングやKDD（Knowledge Discovery in Database）処理等の別の技術を使って行うというものもある。しかしながら、現状“唯一の”評価法として認められている手法は存在していないのが実情である。

この解決を図ろうとしている多くの研究者は、日頃Web管理者が直面しているこれらの問題のいくつかを解決すための方法と手法を開発している。これらの手法は一般にWM（Web Mining）という名前で呼ばれており、このWMはWebシステム（Webページ、Webログ、Webプロファイル等）上のデータに対するデータマイニング技術の一応用と位置づけられている。しかしながらWM分野における研究はより良い検索結果を得るためにまだまだ多くの解決すべき課題が残されている。

また、移ろいやすいユーザの要求を企業側が満足させるためにはこれらすべての改善が迅速かつ簡単に行われなければならない。これは企業間競争に勝ち、企業の生き残るための至上命題であると言える。

本論文では、以上のような背景のもと、WM技術とKDD技術を用いてWebサイトの構造をならびにコンテンツを向上させるための最新技術について研究開発することを目的としている。本論文では特にWebサイトの構造、内容、構成について劇的な変化をもたらすオフライン処理に焦点を当てたものである。

Abstract of Dissertation

Title A Study on Web Mining Techniques for Off-Line Enhancements of Web Sites
(ウェブサイトオフライン改善のためのウェブマイニング技術に関する研究)

氏　名　リオス セバスチァン アレハンドロ
Rios Sebastian A.

(本文)

World Wide Web (WWW) is one of the most remarkable achievements for humanity. Nowadays it is almost impossible to image business or entertainment services without it. Enterprises changed their way of thinking from having a web site just as informative on-line advertisement or just to have one site because it is fashionable; to realizing that it is a new, massive and simple way of doing businesses.

Every day many people are connecting to Internet and probably some of them are active part of the WWW by creating their own contents (by using blogs, Vlogs, own their personal web site[1], etc.) Probably all of them are using a search engine and some may use the WWW to buy products or services, rent a movie from the living of their houses, perform video conferences using a video conference client (i.e OpenWengo, Ekiga from the open source community), or to communicate with their families or friends at very low rates or for free. The amount of services over the Internet has experienced explosive growth as the amount of Internet users rises.

WWW was created by the combination of Internet (a global network) and HTML (Hyper Text Markup Language)[2]. This very simple idea allows anyone to publish its own documents. At first, programming in simple HTML code and uploading the files to the server; later, just using tools, which help to create web documents (knowing nothing of HTML programming, File Transfer Protocol (FTP), among others). These technologies allowed the massive proliferation of web pages, Web Sites, Web Portals, etc. Although, it also brought a problem: "the high amount of unlabeled and semi structured information which makes extremely hard to find out useful information among web pages". Besides, even if we reach an interesting web site; how to not get lost when browsing it?

Today WWW is a sea of documents, and everybody that tries to reach some information must first search into a huge among of useless information and this situation is worsening everyday. The development of the "Semantic Web" arises as the solution for these problems; however, its implementation is planned for a long-term horizon. This new Web is a solution to allow people discover useful information in an easier way in the ocean of documents. The idea is to include descriptions of each document (Metadata) written in the form of an XML, to allow other pieces of software like Semantic Search Engines to read them and give results, which have more relation with the visitors searching goals.

A challenging question for managers and organizations arises: how to develop a "good" web site? In order to keep old customers interested in the web site and gain new ones. In other words, how to build a site that provides valuable information to

[1] http://www.rios.tv which is my personal site
[2] http://www.w3.org/MarkUp/

the visitors to make them return continuously to the site? This is a key issue from business viewpoint, because it is demonstrated that the cost of gaining or loosing a new customer is higher than the cost of keeping them.

The above question is hard to solve and can be decomposed in sub-questions easier to solve, such as: How to create a visual design that allow a simple access to the information in the site without confusing the visitors? How to give the information, product or services that visitors need? How to improve the structure of the textual content for a easy and unambiguous browsing?

Nielsen discusses several web site usability problems and also establishes that effectiveness or utility of the web site is strongly related with its usability. In addition to Web interface, it is possible to improve the visitors browsing experience enhancing the sites' structure and content. If the visitor doesn't find what he/she needs, in a reasonable small amount of browsed pages (also called small amount of *clicks*), he/she will change to browse another site.

There are several studies that have shown that it is possible to improve the usability of a site by improving the sites' structure and content. The problem is that the evaluation of the solutions is usually subjective. Several different studies that have shown to improve the web site content and structure using different techniques such as web mining or Knowledge Discovery in Databases (KDD) process, however, no unique solution and it doesn't exist "the only" way to evaluate the results.

Many researchers motivated by these questions have developed methods and techniques in order to solve some of the problems that managers usually face. These techniques receive the name of Web Mining (WM), which is the application of Data Mining techniques to the data originated in Web Systems (web pages, web logs, web profiles, etc). However, more work in WM field is still required in order to reach better results.

All these improvements must be performed quickly and easily to allow the enterprises to successfully satisfy the changing visitors' needs in order to maintain the competitiveness and ensure the survival of enterprises.

This thesis aims to research the state of the art on Web Mining (WM) techniques and Knowledge Discovery in Databases (KDD) techniques applied to Web site data to enhance its structure and contents. I focused my work in the off-line process, which allow performing drastic changes in the structure, content and organization of a Web site.

Then I propose a new way of performing the web usage mining process, which allows the expert/analyst of the business to combine its own knowledge into the mining process to obtain better contents and structural modifications of the web site.