VIET NAM NATIONAL UNIVERSITY
COLLEGE OF TECHNOLOGY

LE DIEU THU

ON THE ANALYSIS OF LARGE-SCALE DATASETS TOWARDS ONLINE CONTEXTUAL ADVERTISING

UNDERGRADUATE THESIS

Major: Information Technology

HANOI - 2008

VIET NAM NATIONAL UNIVERSITY
COLLEGE OF TECHNOLOGY

LE DIEU THU

ON THE ANALYSIS OF LARGE-SCALE DATASETS TOWARDS ONLINE CONTEXTUAL ADVERTISING

UNDERGRADUATE THESIS

Major: Information Technology

Supervisor: Assoc. Prof. Dr. Ha Quang Thuy
Co-supervisor: Dr. Phan Xuan Hieu

HANOI - 2008

ABSTRACT

With the rise of the internet, there came the rise of online advertising. It in turn has been playing a growing part in shaping and supporting the development of the Web. In contextual advertising, ad messages are displayed related to the content of the target page. It leads to the problem in information retrieval community: how to select the most matching ad messages given the content of a web page.

While retrieval algorithms, such as determining the similarities by calculating overlapping words, can propose somewhat related ad messages, the problem of contextual matching requires a higher precision. As words can have multiple meanings and there are many unrelated words in a web page, it can lead to the miss-match.

To deal with this problem, we propose another approach to contextual advertising by taking advantage of large scale external datasets. Using a hidden topic analysis model, we add analyzed topics to each web page and ad message. By expanding them with hidden topics, we have decreased their vocabularies' difference and improved the matching quality by taking into account their latent semantic relations. Our framework has been evaluated through a number of experiments. It shows a significant improvement in accuracy over the current retrieval method.

ACKNOWLEDGMENTS

Conducting this first thesis has taught me a lot about beginning scientific research. Not only the knowledge, more importantly, it has encouraged me to step forward on this challenging area.

I must firstly thank Assoc. Prof. Dr. Ha Quang Thuy, who has taught and led me to this field and given me a chance to join into the seminar group "data mining". It is one of my biggest chances that has directed me to this way in higher education.

Giving me many advices and teaching me a lot from the smallest things, Dr. Phan Xuan Hieu is one of my most careful and enthusiastic teacher I can have. I would like to send my gratitude to him for his instruction, willingness and endless encouragement for me to finish this thesis.

I would like to thank BSc. Nguyen Cam Tu, my senior at the college, who has supported me a lot in this thesis. I have learnt many things from her and this work is greatly devoted thanks to her previous work.

I would also want to send my thank to all the members of the seminar group "data mining", especially BSc. Tran Mai Vu for helping me a lot in collecting data; Hoang Minh Hien, Nguyen Minh Tuan for giving me motivation and pleasure during the time.

My deepest thank is sent to my family, my parents, my two sisters, their families - my deepest and biggest motivation everlastingly.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRREVIATIONS

CPA	Cost Per Action/Acquisition
CPC	Cost Per Click
CPM	Cost Per Mille/Thousand
CTR	Cost Through Rate
IDF	Inverse Document Frequencies
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
PLSA	Probabilistic Latent Semantic Analysis
PLSI	Probabilistic Latent Semantic Indexing
PPC	Pay Per Click
TF	Term Frequencies

Introduction

"Advertising is the life of trade"[1]. The power of it has grown largely over the past twenty years; and companies are now realizing the potential of the internet for advertising. It is definitely a gold mine and one of the best places for advertising campaigns to start on.

An unfailing question of advertisers over the years is "how to deliver the right advertising message to the right person at the right time?". Target audience in any advertisement is an essential factor because advertising at the wrong group would be a waste of time. With internet, contextual advertising is one of the non-intrusive solutions for this question. Ad messages in contextual advertising are delivered based on the content of the web page that users are surfing, thus increase the likelihood of clicking on the ads. In order to suggest the "right" ad messages, contextual matching and ranking techniques are needed to be used.

This thesis presents an investigation into the problem of matching in contextual advertising. In particular, the main objectives of the thesis are:

- To give an insight into online advertising, its architecture, payment methods, some well-known contextual advertising system like google; and examine the principles to increase its effect to attract customers, with main focus on contextual advertising.
- To learn about online advertising in Vietnam and point out the emergence of an online advertising network; thus predict the potential and applicability of contextual advertising in Vietnam for the next few years.
- To investigate the problem of matching and ranking in contextual advertising, study literature techniques that have been published recently to solve the problem.
- To propose another approach to this problem using hidden topic analysis of a large scale external dataset, then evaluate the performance of this proposed framework through a number of experiments.

The thesis is organized as follows:

[1] Calvin Coolidge, quoted in "*The International Dictionary of Thoughts*", American 30th President of the United States

Chapter 1 provides a general overview of online advertising, its brief history, growth and payment method. We then focus on contextual advertising, a kind of online advertising that its efficiency has been proved through some well-known examples, such as google adsense. We also present some related works on matching and ranking techniques recently, and introduce the challenges to the research community in the field. Chapter concludes by our key ideas, approach and main contribution to the problems using hidden topic models for contextual advertising.

Chapter 2 focuses on online advertising market in Vietnam in order to point out its potential and predict its fast growth and changes in the next few years.

Chapter 3 introduces our general framework for contextual advertising using hidden topic analysis of a large scale Vietnamese dataset in details and explains main advantages of the framework.

Chapter 4 accounts for hidden topic analysis of a Vietnamese collection. We first review the theory and background of hidden topic analysis, with focus on Latent Dirichlet Allocation and Gibbs Sampling method. We then describe our work of hidden topic analysis of a large scale Vietnamese dataset: VnExpress, and its result.

Chapter 5 presents our experiments to evaluate the performance of our proposed framework presented in chapter 3 and discuss the results.

Chapter 6 sums up our main contribution, achievements, remaining issues and future works.

Chapter 1: Online Advertising

Online Advertising is a kind of advertising that use the Internet in order to deliver massages and attract customers. The environment in which the advertising is carried out can be various, like via Web sites, emails, ads supported software, etc. Since its 1994 birth, online advertising has grown quickly and become more diverse in both its appearance and the way it attracts users' attention. One major trend of online advertising that its efficiency has been proved recently is contextual advertising. It is the kind of advertising, in which the advertisements are selected based on the content displayed by users. Its matching techniques have attracted studies and controversies in information retrieval community recently.

This chapter gives an insight into foundations, chronological development of online advertising in the market, its categories and payment methods. In the second section, we focus on contextual advertising, its basic concepts, examples of real-world ad systems, related studies on matching and ranking techniques towards contextual advertising and introduce the challenges to the research community in the field. Chapter concludes by our key ideas and approach to the problems using hidden topic models for contextual advertising.

1.1. Online Advertising: An Overview

1.1.1. Growth and Market Share

In 1994, Internet Advertising began when the first commercial web browser, Netscape Navigator 1.0, released the first banner advertisement [14] . The first ads on the web were static printed ads or company logos. Those banners first appeared at the top of the page because that is where advertisers thought they could get the most visibility.

As technology has expanded to create more opportunities, many new types of online advertising have been developed. Some companies advertised through web-sites by pop-up ads, such as DoubleClick, AdForce and Windwire. They provide some graphic information and tell the browser what to do if a user clicks on an ad [14] .

Web sites, which are driven by databases, give a new dynamic interaction that allows sites to deliver information based on a user's input. Most often, this user-specific information is stored on the user's computer in the form of "cookie", which helps the

Web browsers to remember the user's identity. To take advantage of this information, many systems have analyzed it to provide recommendations of merchandise that the user should be interested in based on his preferences or even past purchases. One well-known example of such system is Amazon.com.

The new decade of engine technologies created a new level of online advertising [21]. A successful advertising system that based on search engine is Google AdWords, which allows advertisers to display advertisements in Google's search result. In 2005, Google announced a beta version of AdSense system. According to the Officer Google Blog, advertisers can place their ads in most appropriate place and readers can see relevant things with this system.

A decade after its first appearance, advertiser in the U.S. market spent $9.6 billion on Internet ads, grew at the rate of 31.5% from 2003 to 2004 [21]; compared to 10% for broadcast TV, 7.4% for the advertising industry in general (Universal McCann) and 6.6% for the current-dollar GDP of the U.S. economy (Figure 1). According to the report of IAB in February 2008, Internet Advertising revenues have reached new highs, estimated to pass $21 billion in 2007.

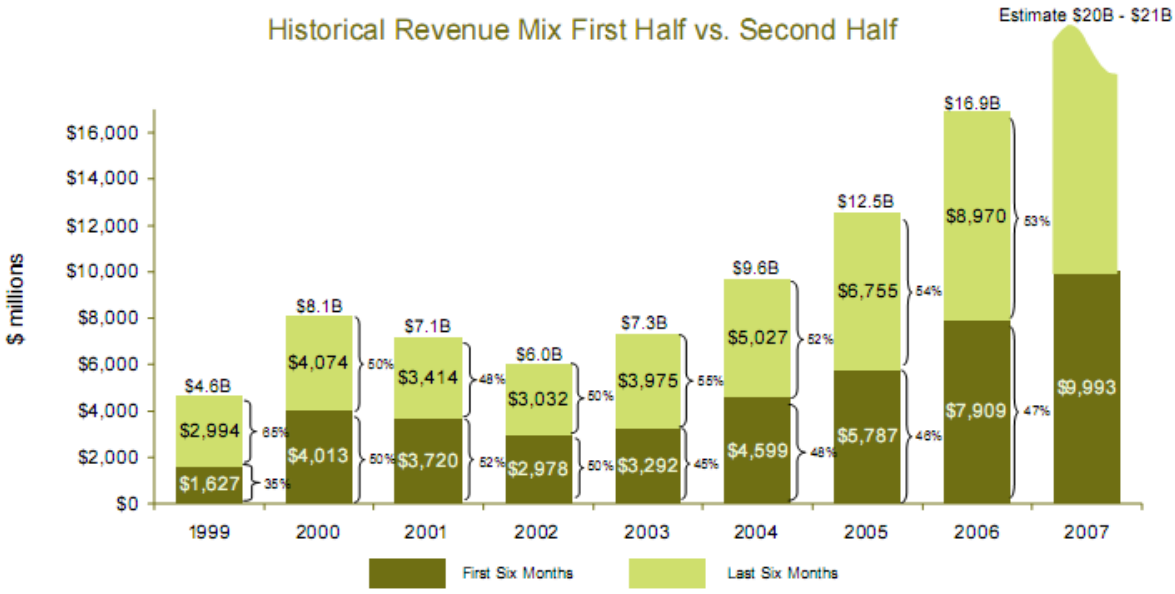


Figure 1. Online Advertising Revenue Mix First Half versus Second Half from 1999 to 2007 in the U.S.

According to the latest report by Strategy Analytics [29], global expenditure on online advertising rose by nearly a third to $47.5 billion in 2007 and is set to pass the $100 billion mark by 2012.

This brief history of online advertising and its steadily growing revenue promise that online advertising will continue to change and grow in a fight for the future.

1.1.2. Advertising Categories

Online Advertising can be categorized as legitimate (advertising networks) and illegitimate (spamming).

The spamming advertising is often intrusive that is usually labeled as Spyware, Adware or Pop-up advertisements. For example, when a new browser is opened, pop-up ads appear to drive traffic to the sponsor's websites. Because of their annoyance, many browsers provide pop-up blocking feature to restrict illegitimate pop-ups. Spyware and Adware are often external applications. Some of them are really harmful, like Trojan.

Legal advertisement can be classified into Display Advertising, E-mail, Classifieds/Auctions, Lead Generation, Rich Media and Search, which distributed revenues in first six months of 2006 and 2007 in U.S. are illustrated in Figure 2 [19].

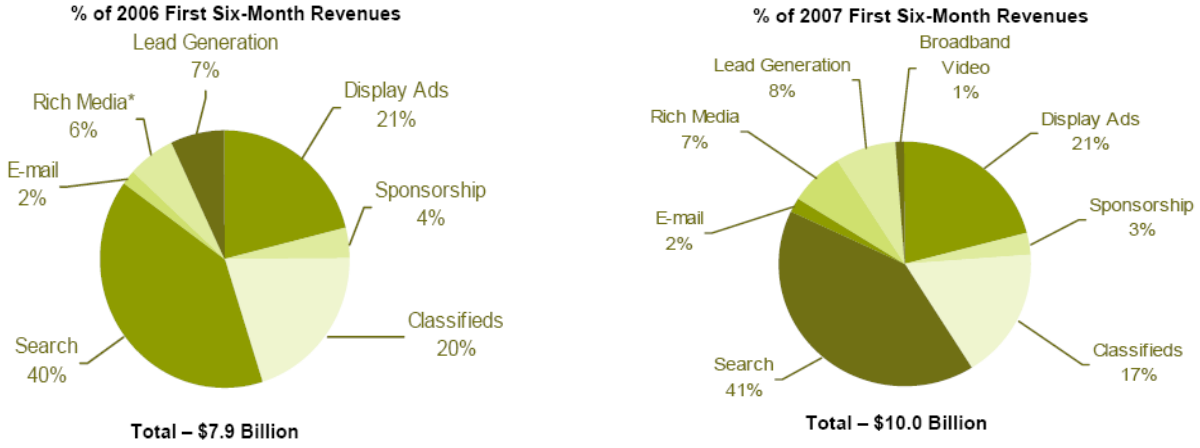


Figure 2. Online Advertising Revenues by Advertising Categories in first six months in 2006 and 2007 in the U.S.

Display Advertising is often placed as a static or hyperlink banner or logo on an Internet company's pages and advertiser pays the company for the space. For the first six month of 2006 of 2007, it holds 21 percent of total revenues.

Sponsorship advertising generally occurs when an advertiser pays to advertise on all or some sections of a website, which content is related to but not competitive with the services provided by the sponsoring company. Normally, it can take the form of traditional banners with sponsored content like "sponsored by". Its revenues accounted for 3 percent, down slightly from 4 percent reported for the same period in 2006.

Email is another kind of online advertising, in which links or advertiser sponsorship's content are delivered through newsletters, accounted for 2 percent of total revenue.

Lead Generation is the fee advertisers pay to Internet advertising company when they provide consumer information like contact, behavior, survey, contest, etc. Its revenue was up slightly from 7 to 8 percent for the same period.

Classifieds and auctions are fee advertisers pay to Internet companies to list and categorize items and products like yellow page or real estate listings.

Rich media is now becoming more attractive to advertisers as it can help marketers reach customers interactively using animation, sound or video. Flash, Real Video/Audio, Shockwave, applets and other technologies allow new level of advertisement, which is more colorful and animated.

Broadband Video Commercials are TV-like advertisements. They appear in a streaming video, animation, gaming or music video content.

Search advertisement refers to placing ads related to a domain by a specific search word or phrase. It includes paid listings, paid inclusion, site optimization and contextual search. Paid listings are text links that appear on one side of search results, corresponding to specific keywords. Their positions are determined by the payment of advertisers.

Paid inclusion ensures that advertisers' websites are indexed by search engines while site optimization makes it more possible for a website to be listed in search results by modifying the site.

Contextual search or contextual advertising is text or other kinds of link that is chosen to be appeared based on the context of the content. The payment is made when the link is clicked or some actions occur. The payment methods will be discussed in more detail in the next section.

As can be seen in figure 2, search advertising including contextual search remains the largest revenue type of internet advertising in the U.S. market from 2006 to 2007 and has been increased steadily. It accounts for 41 percent of total revenue coming from internet advertising in the first six months of 2007.

In summary, there are many kinds of online advertising, which can be categorized as legal (Display Advertising, E-mail, Classifieds/Auctions, Lead Generation, Rich Media and Search) and illegal advertising (spamming, Adware, Spyware). In legitimate category, search advertising including contextual advertising has become the most popular and brought the largest revenue to the internet advertising market according to the report of Price Water House Coopers last year in the U.S.

1.1.3. Payment Methods

There are three common ways in which online advertising is paid: CPC (Cost Per Click) or PPC (Pay Per Click), CPA (Cost Per Action or Cost Per Acquisition) and CPM (Cost Per Mille – thousand).

In CPC model, the advertisers pay for every time their link is clicked. Although it is not a good indicator of whether or not there is any real impact of the advertisement to the advertisers' company, it is still widely used.

CPA model answers to the question in CPC model; the payment is only made when a user completes a transaction, such as a purchase or sign up. It helps advertisers discover how much it costs on the Web to acquire a new customer.

While CPA gives advertisers a specific payment by a performance based method, CPM appears to be the most imprecise model. It is where advertisers pay for exposure of their message to a specific audience. It estimates the cost per 1000 views of the advertisement. For example, if a website sells banner ads for a $20 CPM that means it costs $20 to show the banner on 1000 page views. This model is often used in marketing to calculate the cost of an advertising company, normally ranges from $10 to $30 CPM.

While those three models of payment help the advertisers to estimate their profits, CTR (Click Through Rate) measures the success of an online advertising company. It defines the number of users who click on an ad on a web page by the number of times the ad was delivered. For example, in 100 times the ad appears on a web page, one user clicks

on the ad, it can be concluded that CTR is 1 percent. The task of online advertising company is trying to maximize the number of CTR by improving the impression to users and then to increase their benefits as the result.

1.2. Online Contextual Advertising

As mentioned above, contextual advertising is a kind of online advertising, which ads are chosen to display depending on the content of a web page. It can be categorized to search advertising group, which revenue accounted for 41 percent of total revenue coming from online advertising in the U.S. in the first six months in 2007.

This section focuses on contextual advertising model, its basic concepts and introduces contextual matching and ranking techniques that have been proposed for this advertising model recently.

1.2.1. Advertising Network

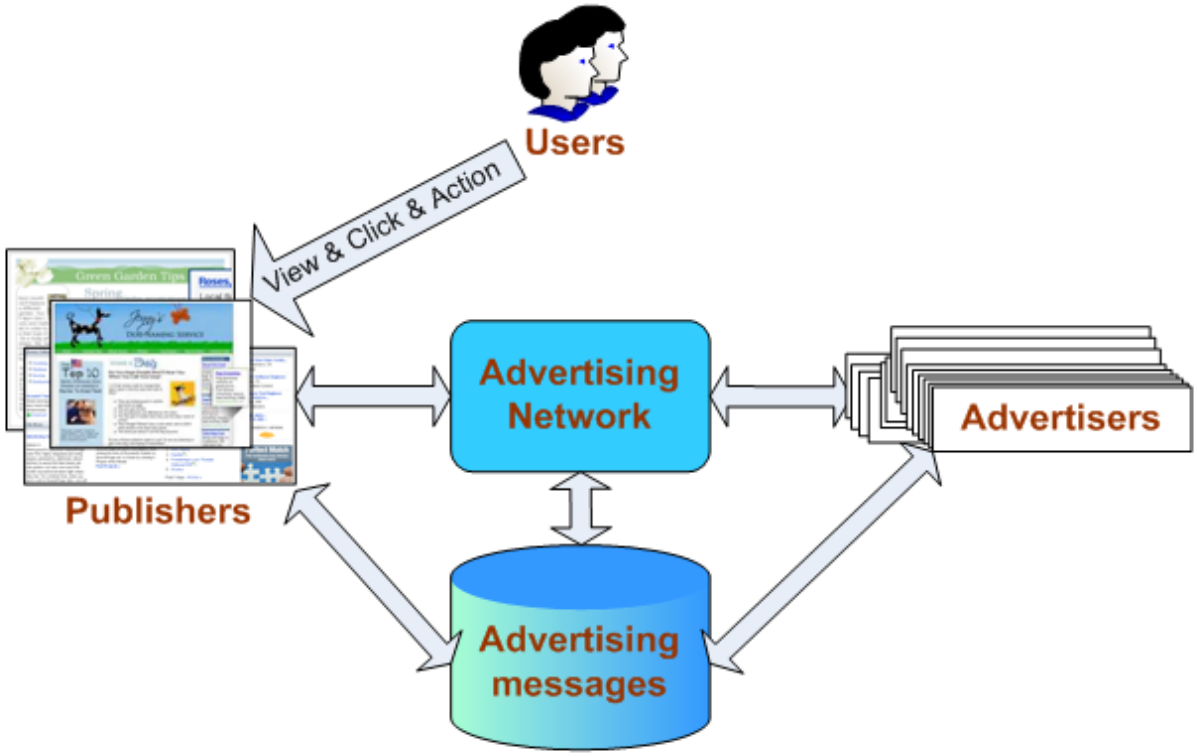


Figure 3. Online Contextual Advertising Architecture

While Sponsored search ads are placed beside a search's result related to the query of the user, contextual ads are displayed in a web page, which content is relevant. Figure 3 illustrates the architecture of an online advertising system.

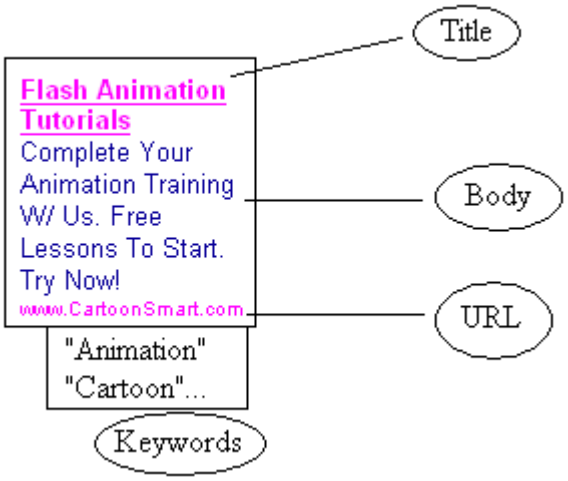


Figure 4. An advertising message form

Through an advertising network, ad messages are delivered to different web pages of publishers based on their contents. When a user clicks or takes some actions, advertising network will recognize and the advertisers will pay for the click or action depending on the business model. The revenue will be shared between publisher and advertising network (Figure 3).

Advertising message normally can be composed of four parts: title, body (description), URL and bid-phrases (or keywords). They are often used to evaluate the relevance to the content of the displayed web pages (Figure 4).

Figure 5. Google AdSense example

Google AdSense (Figure 5) is an example of the advertising network.

Most of the revenue of Google comes from advertising. These days, we can see google's ads on many web sites and it can be considered as the first truly successful contextual advertising service.

Other examples of such networks are Yahoo! Publisher Network (YPN); eBay AdContext; Amazon.com, providing suggestion Book Ads; MIVA Monetization Center with three services for web publisher (Content Ads, MIVA InLine Ads and Search Ads); Clicksor.com, etc.

1.2.2. Contextual Matching & Ranking – Related Works

The main task of a contextual advertising model is to decide which ad messages to display given a targeted page and a set of ads. It introduces new challenging technical problems and raises the question of how to match and rank the ad messages given the content of a webpage.

Different from sponsored search, which ad messages are chosen depending on only the keywords provided by users, contextual ads depends on whole content of a webpage. Keywords given by users are often condensed and reveal directly the content of the users' concerns, which makes it easier to understand. Analyzing web pages to capture the relevance is a more complicated task. However, contextual matching is a more potential area for providers as the time users spend on web pages is much more in compared with search pages. Recently, there have been a lot of studies and controversies around this area. Example of these studies includes keyword extraction strategies [38], semantic approaches [12], impedance coupling [13] and ranking optimization [11] that will be discussed in more details hereafter.

- **Keyword-based models**

Originated from the idea of sponsored search, we can consider targeted page as a long query or extract keywords from the page. Yih et al (2006) [38] has proposed a supervised system that can extract keywords for advertising target. Training from a set of pages that have been keyword-defined, they use a classifier using machine learning with logistic regression learning algorithm.

To determine which keywords or key phrases that best describe a web page, they used several methods for selecting and carried out experiments to find out which method had the best performance. They considered three methods: MoS, MoC and DeS. M (Monolithic) means considering the whole phrase as a candidate. D (Decomposed) considers each word in a phrase as a distinct one. S (Separate) means that different words or phrases even with the same content will be regarded as different candidates, whereas C (Combined) will combine same words/phrases as one.

One important point of their work is that they use 7.5 million queries from query logs of MSN [24] as a feature for selecting, together with 11 other features, such as information retrieval oriented feature (term and document frequencies), linguistic feature (using pos tagging), capitalization (whether a word is capitalized or not), hypertext (whether a candidate is an anchor text or not), title (HTML header of a page), phrase or sentence and document length, etc.

In their experiments, they used a set of 828 web pages chosen from Internet Archive [20] to train and test the system. It shows that the MoC selector, in which identical phrases are combined as one, performs the best result whereas the separate MoS system is the worst. In addition, the DeS system that considers each words as separately is significantly worse than the monolithic approach that consider whole phrases. The accuracy of the best one is 30.06% in compared with 13.01% of a simple model using TF-IDF.

To learn the contribution of each feature, they conducted experiments in the same system removing and adding each feature in turn. The result points out that query logs and IR feature play the most important part as it affects the score most significantly.

Their study provides an approach to contextual advertising problem inspired by the query-based ranking problem, which has been better understood. Their framework allows ranking the ads based on extracted keywords from web pages. However, the relevance of chosen ads based on extracted keywords in this system has not been proved through experiments yet.

- **Semantic Approaches**

While extracting keywords from web pages in order to compute the similarity with ads is still controversial, Andrei Broder at al [12] proposed a framework for matching ads based on both semantic and syntactic features.

For semantic feature, they classify both web pages and ads into a same large taxonomy with 6000 nodes. Each node contains a set of queries. They carried out three experiments with three different classifiers: SVM, log-regression classifiers and a nearest neighbor classifier. With the first two classifiers, they prepare a training set by running the given queries over a web search for training pages and selecting ads for each class based on keywords. The third classifier, which uses only those queries as centroids for each group, is the best among them. It is probably due to the robustness of the training set using search engine.

For syntactic feature, they used the tf-idf score and section score for each term of web pages or ads. The section score can be determined based on the importance of each section (title, body or bid phrase section).

To compute the similarity of a page and an ad, they introduced a function that is combined of semantic and syntactic score with an external parameter. On evaluation, they use 105 pages and nearly 3000 ads and report an improvement of around 30 percent precision when using both semantic and syntactic feature against using only syntactic one.

- **Impedance Coupling**

One problem of contextual matching task is the difference between web pages and ads' vocabularies. Ribeiro-Neto et al (2005) [13] focuses on solving this problem by expanding the vocabulary of web pages.

Generally, web pages have richer content and belong to a larger contextual scope than an ad. They can be about any subject with many specific terms. However, ad message is often short, condensed and focuses on a main subject with more general terms. Moreover, how we can find good ads for a specific web page when sometimes unimportant topics in the page can offer good opportunities for advertising is still questionable.

In order to solve this problem, Ribeiro-Neto et al (2005) [13] has proposed 10 matching strategies. They conducted an experiment using real case database with over 93,000 ads and 100 Web pages for testing.

For the first five strategies, they matched web pages and ads using standard vector model. The ranking of each ad is computed by the cosine similarity with each page. They match the ads based on their titles and descriptions, their keywords sequentially. The best among those methods is AAK method, which stands for "match the ad keywords and force their appearance in the web page", and will be used for baseline in the impedance coupling method.

As described above, there is often a distinction between the vocabulary in the web pages and that in the ads. To overcome this, they expand the page vocabulary with terms from other similar pages decided by means of a Bayesian model. Those extended terms can be appeared in ad's keywords and potentially improve the overall performance of the framework. For better understanding about the content of these short ads, they also carried out an experiment that considers the page pointed by the ads in advance.

In their experiments, they used a database of about 6 million web pages crawled to generate expansion terms. It shows an increase in the precision against the baseline method. The best strategy of all is the one using expansion terms and also considering the content of the landing pages pointed by the ads.

The experiments of Ribeiro-Neto et al (2005) have proved that when decreasing the vocabulary distinction between web pages and ads, we can find better ads for a targeted page.

- **Ranking Optimization with Genetic Programming**

Following the former study [13], Lacerda et al (2006) [11] introduced a new approach based on Genetic Programming to improve the ranking function. Given the importance of different features, such as term and document frequencies, document length and collection's size, they use machine learning to produce a matching function to optimize the relevance between the targeted page and ads. It was represented as a tree composed of operators and logarithm as nodes and features as leaves. They used a set of data for training and a set for evaluating from the same data set used in [13]. It has shown a better gain over the best method described in [13] of 61.7%.

1.3. Challenges

Online advertising in general and contextual advertising in particular are potential areas of research. They have motivated studies in different fields, but also introduced new challenges. In order to attract customers, we have to find the best matching ads with a targeted web page. The "best matching ads" is also difficult to define as web pages are about different contents with different topics. The challenge is also how to extract the customers' interest from such web pages in a diffuse context. Furthermore, even unimportant topics can offer good opportunities for advertising. For example, a web page about a scientific conference in Hue province should also provide an ad about hotels in Hue, as people who might go there would also consider that information.

Moreover, meeting the requirement of real time application with the huge data and transactions also appears to be an important part of contextual advertising. Hence the systems need to be able to deal in real time to serve people in different languages with a good quality matching algorithm. Another important point of these systems is that the ranking function also needs to balance the importance of high click-through-rate (CTR) with advertiser's willingness to pay. In other words, the ultimate ad messages are chosen taking into account the congruence between ad messages and context of web pages and also the price of the ads.

1.4. Key Idea and Approach

As has been discussed by Ribeiro-Neto et al (2005) [13], there are two key issues with contextual matching and ranking for advertising problem. First, the vocabularies of the targeted page and advertisements are often different as web pages often belong to a broader scope. Second, a good advertisement of a targeted page might pertain to a topic that is not mentioned explicitly in the page. Besides, Broder et al (2007) [12] and Ciaramita et al (2008) [23] have noticed that standard matching approach can be improved by taking into account the semantic relations, such as topical proximity.

Based on the idea that expanding web pages and ads with external terms will offer a better opportunity for finding the right matching ads, we propose an approach to contextual match that focuses on topic analysis and enriching both web pages and ads with external terms. In order to generalize the context of web pages and ads, we first learn the framework with the support of topic model estimated from a large universal

dataset. That will help us to discover the hidden topics and capture the relations between topics and words as well as words and words in our domain, thus partially decrease the limitation of word choices. Through the learning model, we can again analyze the topic distribution of web pages and ads in order to enrich them with hidden topics or new terms of the same topics.

In general, our key idea is based on the fact that matching web pages and ads relied on only their given terms may not provide us a satisfactory result, we can improve the performance by expanding them with topic analysis models like Latent Dirichlet Allocation (LDA). The underlying idea is based on topic analysis of available large scale dataset.

1.5. Main Contribution

Bearing in mind the importance of reaching target audience in advertising, studies [37] have shown that one of the main factors of a success contextual advertisement is their relevance to the surrounding context. Finding the most relevant ad messages has been an emergent field of study though public literature in this field is still very sparse. A nature matching using retrieval information such as counting words overlap is insufficient. As words can have multiple meanings and some words in the targeted web page are not important, it sometimes leads to miss-matches.

To deal with this problem, we have proposed another approach that can produce high quality match that takes advantages of external large scale datasets, which are not "expensive" and easy to collect in the internet. Our framework is also easy to implement and general enough to be applied in different domains of advertising, different languages.

Through a number of experiments, it also indicates that this framework can suggest appropriate ad messages for contextual advertising and can be practical in reality.

1.6. Chapter Summary

This chapter brought an overview of online advertising in general and contextual advertising in particular. After introducing its architecture, payment method, we then focus on the major problem in contextual matching and ranking. Some remarkable issues related to this diminished problem were introduced in section 1.2.2. We reviewed four studies including keyword extraction strategies, semantic approaches, impedance

coupling and ranking optimization, which have been proposed recently. After examining the problem with related works, we introduced the challenges, then propose another approach using hidden topic analysis and summarize our main contribution through out this thesis.

Chapter 2: Online Advertising in Vietnam

We have introduced about online advertising and its widely applicability and potential in many countries. In this chapter, we will provide an overview of online advertising in Vietnam, thus predict its fast growth and point out the necessary emerge of an online advertising network in the next few years.

2.1. An Overview

2.1.1. Market Share

As the internet computer market grows rapidly, Vietnam's online advertising potential is at its first great peak. A country of more than 80 million inhabitants with the GDP (Gross Domestic Product) growing by 7.5 percent annually is a good business environment. Vietnam is currently a fledgling market for online advertisement, but it has a lot potential [4] .

Figure 6. Online advertising in a Vietnamese e-newspaper (May, 2008)

The online advertisement revenue in Vietnam is estimated to be 160 billion VND in 2007 and predicted to increase by 100 percent to reach 500 billion by 2010 [6] Though expected to grow at a very fast rate, it is still very new and quite unfamiliar with advertisers up to now. Currently, 80% of domestic advertisement belongs to broadcast on television and the second market share is advertisements on newspapers. However, online advertisement holds only 1.3% of total advertisement revenue in Vietnam [6] .

Still in its infancy but potential, it is high time Vietnam advertising market took into account online advertising in order to expand their revenue and improve enterprises' advertising campaign.

2.1.2. Advertising Categories

At present, online advertising's categories in Vietnam fall into some common groups, such as banner, pop-up, in-line, newsletter and multimedia advertisements. All of those are often placed in high ranking e-newspapers with a large number and in confusion with many colors (Figure 6). That makes it difficult and annoying for visitors to follow (according to Laodong e-newspaper). Moreover, advertisements are displayed not in any order, subjects or selection. Targeted and contextual advertising are still new concepts for advertisers and publishers. No strategy for selecting appropriate advertisements is applied. Additionally, most of the advertisements are lying on some high ranking e-newspapers such as VnExpress, DanTri, VietnamNet, etc. but have not taken the advantage of a numerous domain web site about particular subjects like travel, food, medicine to advertise to a specific kind of audiences.

Still keeping in mind the payment method of traditional advertising in printed newspapers, publishers and advertisers in online advertising are contracting using the price calculated by sizes of banners and the number of exposition through the ranking of publishing web sites (CPM method). This ranking is often provided by some tools adopted in the internet, e.g. alexa.com. The price is decided based on the number of visitors to the website and the position of the banner.

Other payment methods like CPC or CPA are still very rare as there has been a need of a trusted advertising network that can provide statistics of traffic ranking to support the framework. This is also an important issue that explains why contextual advertising in Vietnam has not yet been developed. However, some active companies have caught this trend and are testing the new framework with CPC payment method, such as Hura ad[2], daugia 247 – ECOM JSC[3] and VietAd[4], which system had once been tested in VietnamNet websites (but has been removed to improve by now, according to VietnamNet).

CPA payment method (that payment is made only when users complete some actions before clicking into the landing page like purchase) has not yet been considered

[2] http://ad.hurahost.com
[3] http://daugia247.com
[4] http://vietad.vn

here as it requires a more developed e-commerce, which will be discussed in more details in section 2.2.1.

In general, online advertising market in Vietnam has few players and few forms or types. It is at the beginning period. Advertisements are often banners and placed statically in a website and paid based on its size or position and on the ranking of this website.

2.2. Untapped Resources and Markets

In the previous section, we have introduced a general view of the infancy but opening and potential online advertising market in Vietnam. In this section, we will explain more in detail the untapped resources and markets to point out the potentiality and the emergence of an online advertising network in Vietnam in the next few years.

2.2.1. Rapidly Growing E-Commerce System

As mentioned above, e-commerce is an important factor of online advertising, especially for the payment method of a targeted and contextual advertising system. When e-commerce develops, more business can take the advantage of trading through the internet. That will be a fertility land for online advertising to cultivate. In other words, e-commerce growth will provide a framework for small mass markets to introduce their products to customers and that will support the development of contextual advertising as a result. If well-known brand names are now considering online advertising as a minor choice for their advertising campaign, it will be acceptable to advertise through traditional banners only. However, the success of contextual advertising in other developed countries has shown that not only well-known brand names but also mass markets are potential field of online advertising. Online advertising is cheaper and more convenient, so it will be a major choice for many mass markets.

In brief, e-commerce will encourage not only big but also small businesses to develop their websites and trade through the internet. Online advertising will thus provide major income for e-newspapers, online companies and also bring money to all the online communities. Contextual advertising will become an important type of advertising consequently.

In June 2006, e-commerce began to take shape and new decree-laws were promulgated. With the support of government, e-commerce in Vietnam has made great advances and is believed to impulse the development of the economy [2].

2.2.2. Explosion of Online Communities and Social Networks

Recently, there has been a new trend of using the world wide web technology and web design that make it easier for users to share their own information, such as social-networking sites, wikis, blogs and forum. It can be called Web 2.0. In line with this new trend, the number of Vietnamese Internet users is increasing considerably these years and has created big online communities and social networks among Vietnamese users. According to VNNIC (Vietnam Internet Association), in March 2008, the Internet users in Vietnam has reached over 19 million (19.41 percent) and is growing at a potential rate. The market is bigger than that of Thailand, Philippines and Indonesia. Over the past few years, the online communities have experienced the development and fierce competition of social networking sites, both from local and overseas co-operations, such names as Yahoo! 360 blog, Tamtay, Yobanbe, Cyworld, Zoomban, etc.

Of course, there seems to be a gap between the development of e-commerce in Vietnam and that of other developed countries as it partially depends on the users' habit and income. However, since internet users are getting acquaintance with internet shopping and advertising, Vietnam is definitely a rising potential market.

2.2.3. Proliferation of News Agencies and Web Portals

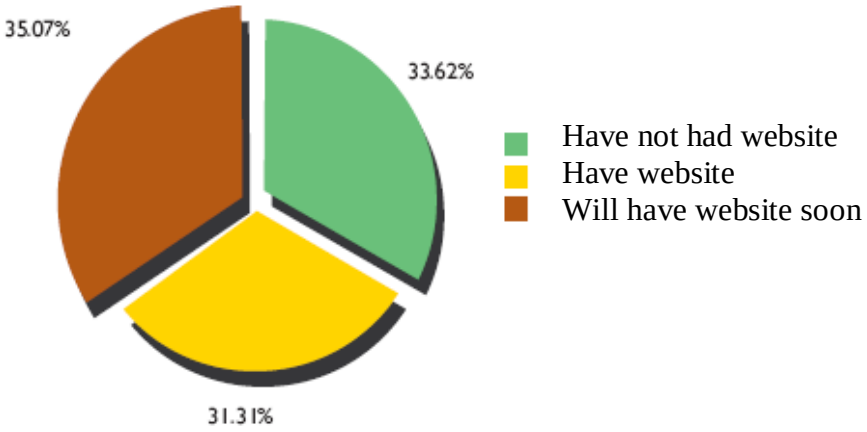


Figure 7. The percentage of companies having website, not having website and will have website soon (according to a survey on 1,077 businesses by the Department of Trade, 2007)

Along with the growth of online communities and social networks, more and more news agencies and web portals were constructed in order to seek users and monetization. According to the survey carried by the department of Trade on 1,077 businesses last year, the number of those that had their own websites is 31.3 percent and those that will have website soon is 35.07 percent (Figure 7).

Besides, there are more and more Vietnamese e-newspapers built on the internet that attract a large number of visitors, such names as VnExpress, VietnamNet, DanTri, etc (Table 1). Those websites are providing online advertising services and gaining gradually revenue.

Một số website lớn cung cấp dịch vụ quảng cáo trực tuyến

STT	Tên	Địa chỉ
1	Báo điện tử Vnexpress	http://vnexpress.net
2	Báo điện tử VietnamNet	www.vnn.vn
3	Báo điện tử Thanh Niên	www.thanhnien.com.vn
4	Báo điện tử Dân trí	www.dantri.com
5	Báo điện tử Lao động	www.laodong.com.vn
6	Báo điện tử Vn Media	www.vnmedia.com.vn
7	Ngôi sao	http://ngoisao.net
8	Công ty Cổ phần Quảng cáo dịch vụ trực tuyến	www.24h.com.vn
9	Công ty Truyền thông đa phương tiện (VTC)	www.vtc.com.vn

Table 1. Some high ranking Vietnamese websites provides online advertising [2]

2.3. Emergence of Advertising Networks: A Long-term Vision

The rapidly growing E-commerce system, the explosion of online communities and web portals of Vietnam have made a stable foundation for online advertising to develop. It will definitely become a fertile area for local and overseas businesses to exploit.

Recently, Vietnamese internet users have witnessed the advertising campaign of Google and Yahoo in this market. Realizing the potential growth of Vietnamese online advertising, they are preparing for a new marketing strategy and building different services for Vietnamese users. According to VietnamNet, Google is now mobilizing volunteers to translate their services to Vietnamese, such as their adword advertising service[5]. Yahoo is holding the upper hand for having the largest number of users (according to the ranking from alexa). They have just released Vietnamese yahoo version[6] and the new version of blog 360 plus in order to attract users in this market. Their advertisements of new services are broadcasted on Vietnamese television from May this year.

However, the online advertising market has attracted not only overseas but also local companies. Some new and creative companies started to expand their business area to marketing and aimed at online advertising. Vietnamese users have got acquaintance with some high ranking e-newspapers, such names as VnExpress and VietnamNet. Their revenues from online advertising have increased regularly (figure 8) and VnExpress still holds the first place in online advertising on e-newspapers market.

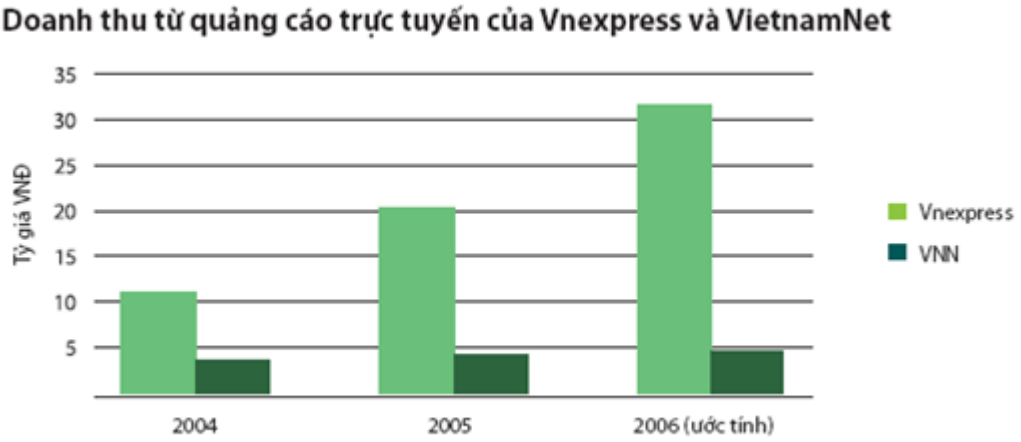


Figure 8. Online Advertising Revenue of VnExpress and VietnamNet e-newspapers [2]

In summary, online advertising market in Vietnam is still in an early stage of development and, as a comparison of VietnamNet, a "new cake" for both local and

[5] http://adwords.google.com/select/?hl=vi
[6] http://vn.yahoo.com/

overseas companies to share. There has been a need of an online advertising network in Vietnam and it is high time new types of online advertising such as contextual advertising became popular.

Google and Yahoo have succeeded in overseas markets. However, the barriers of language and culture made it difficult for them to predominate over all the market in Vietnam. A lesson from the success of Baidu (the leading website of search engine in China) has shown that overseas companies like Google and Yahoo do not always succeed in local markets, especially in Asia [3] . Vietnamese users are still waiting for a Vietnamese network from local companies. Building and developing online advertising networks have become an essential requirement in a long term vision and Vietnamese users will soon experience the fast growth and changes in the advertising market in the next few years.

Chapter 3: Contextual Matching/Advertising with Hidden Topics: A General Framework

In section 1.4, we have introduced our key idea and approach based on two important issues: First, there is often a difference between the vocabulary of web pages and ads that make it difficult for matching. This vocabulary impedance can be solved by expanding web pages with external terms [13]. Second, individual phrases and words might have multiple meanings that unrelated to the overall topic of the page and can lead to miss-matched ads. Therefore, semantic relation is an important factor of a successful advertising system [12] [23]. Inspired by these ideas, we propose a framework for contextual advertising based on the analysis of a large scale dataset as follows (Figure 9).

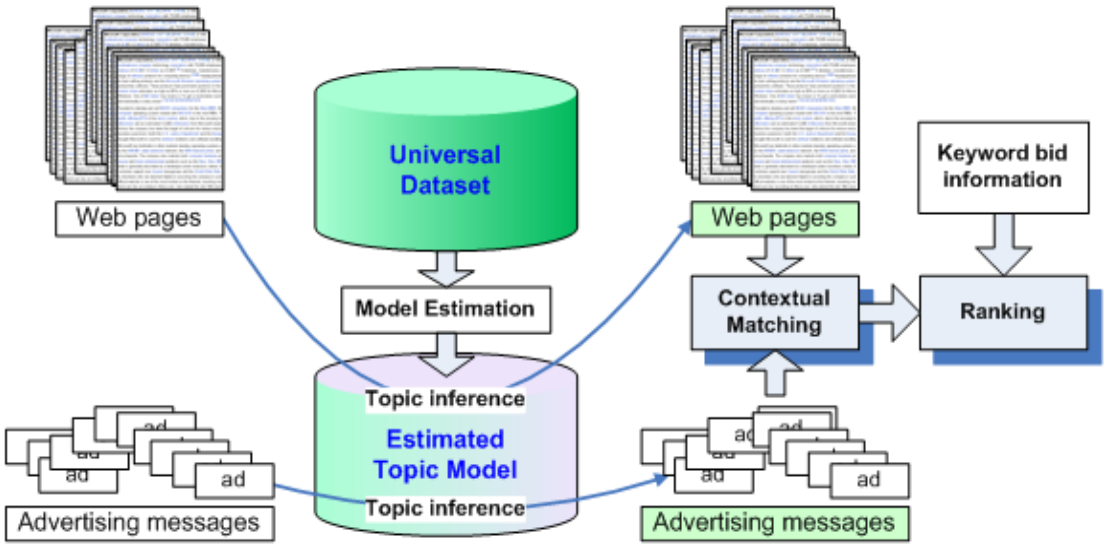


Figure 9. Contextual Advertising general framework

(1) Choosing an appropriate "universal dataset"
(2) Doing topic analysis for the universal dataset
(3) Doing topic inference for web pages and ad messages
(4) Matching web pages and ad messages
(5) Ranking ad messages to the corresponding web page

3.1. Main Components and Concepts

The problem we focus on is that given a web page and a set of advertising messages, matching and ranking them depends on their relevance to the content of the targeted web page. The problem is defined as follows:

Given a set of n pages $P = \{p_1, p_2, \ldots, p_n\}$, and a set of m ad messages $A = \{a_1, a_2, \ldots, a_m\}$.

For each web page p_i, we have to find a corresponding ad message rank list: $A_i = \{a_{i1}, a_{i2}, a_{i3}, \ldots, a_{im}\}$, $i \in \{1..n\}$, such that more relevant ads will be placed before less ones.

As illustrated in Figure 9, first, (1) we collect a large scale dataset for hidden topic analysis. It is based on the idea of modeling text corpora in order to find short descriptions of the members of a collection while preserving the essential statistical relationships [15]. The short description here is the probability distribution of a document over topics and distribution of a topic over terms. After discovering these distributions and hidden topics, we can use them to enhance the matching performance. In general, the result of the step (2) is an estimated topic model that includes hidden topics discovered through the dataset and the distributions of topics over terms.

After the estimating process (2), we can again do topic analysis for both web pages and ads based on this model in order to discover their meaning and topic focus (3). With the distributions of documents over topics that have been estimated in the previous step, we can then add new topic names to our web pages and ads based on their topic distribution. After the combining process, they will be called “new web pages” and “new ads”. Those new web pages and ads, which have been enriched with hidden topics, will be matched using a cosine similarity based on term frequencies (4). The ultimate ranking function can also be adjusted based on its keyword bid information. Ad messages, which keywords given by advertisers will be ranked according to the relevance with the web pages and the money the advertisers pay for them (5).

In the scope of our work, we only focus on the task of ranking based on ads’ relevance and do not take into account the keywords bid information. Hereafter, we will discuss further the process of each component in our framework.

3.2. Universal Dataset

The first important thing to consider in this framework is choosing an appropriate large scale dataset, which is so-called Universal Dataset. Motivated by the idea of exploiting available large datasets, we use this dataset for topic analysis and then enrich both web pages and ad messages with topics extracting from that. In order to take the best advantage of this Universal Dataset, we need to find an appropriate data for our web pages and ad messages. Firstly, it must be large enough to cover words, topics and concepts in the domains of our web pages and ads. Secondly, the vocabularies of the Universal Dataset must be consistent with that of web pages and ads, so that it will make sure topics analyzed from this data can overcome the vocabulary impedance of web pages and ads. The Universal Dataset should also be pre-processed to get a good result. In order to take best use of this dataset, we should remove noise and non-relevant words to enhance the performance of topic analysis process.

3.3. Hidden Topic Analysis and Inference

After choosing and preparing a suitable Universal Dataset for web pages and ad messages, the next step is applying a topic analysis model to this dataset.

Topic models are based upon the idea that documents are composed of different topics, each topic in turns is a probability distribution over words. It can be modeled as a process of generating new documents. The underlying idea is as follow: To make a new document, we can firstly choose a topic distribution for this document. After that, random topics will be chosen according to this distribution and then, words will be obtained from each topic. Consequently, the document has been generated.

The reverse of this process is inference. We can use different standard statistical method to do the inference. That means inferring the set of topics that were responsible of generating those documents. The hidden topic analysis will be described more in section 4.1. In general, we can apply some hidden topic analysis models such as pLSI (Hofmann, 1999, 2001) or LDA (Blei at al, 2003) [15].

In this framework, we use topic analysis for the universal dataset using LDA, which will be introduced in section 4.1. After performing the model estimation, we can represent the content of words and documents with probabilistic topics. Each topic will have a

distribution over words and therefore represent the coherence of different terms. To exploit this representation, we then do topic inference for both web pages and ad messages. The result of this step is the topic distribution of each web page and ad message. By analyzing their topics, we can add these hidden topics to them before matching, thus decrease the difference of vocabularies between web pages and ads.

3.4. Matching and Ranking

After enriching both web pages and ad messages with hidden topics analyzed from the model, we match them using cosine similarity based on term frequencies. Cosine similarity is a vector-based method that measures the similarity of two given strings.

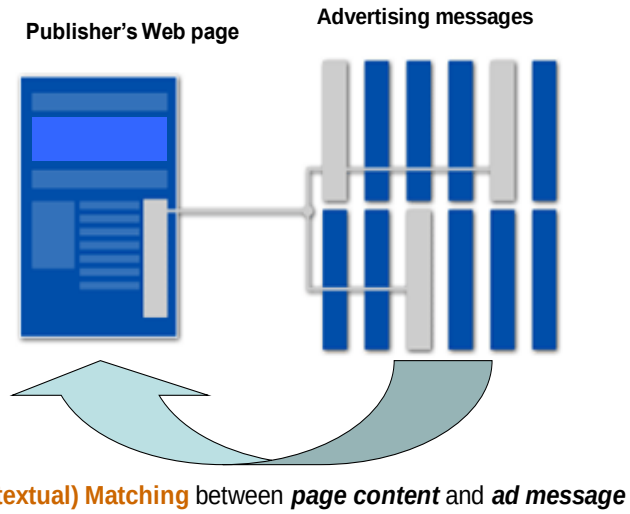


Figure 10: Matching and ranking ad messages based on the content of a targeted page

The basic idea is to represent each string in a vector of some high dimensional space such that similar strings are close to the others. The cosine of the angle between two strings measures the similarity of them. It defines how similar they are.

For a web page p and an ad message a, let w_{pi} be the weight associated with term i in page p and w_{aj} be the weight associated with term j of ad a. Thus, we can represent the term vectors of p and a in a n-dimensional space as:

$$\vec{p} = (w_{p1}, w_{p2}, \ldots, w_{pi}, \ldots w_{pn})$$

$$\vec{a} = (w_{a1}, w_{a2}, \ldots, w_{aj}, \ldots w_{pn})$$

The term specific weight using here is term frequencies: $W_{t,d} = tf_t$, where TF measures the importance of the term within the document.

The cosine similarity of these documents can be calculated with:

$$\text{sim(p,a)} = \frac{\vec{d_p}\vec{d_a}}{|\vec{d_p}||\vec{d_a}|} = \frac{\sum_{t\in T} w_{p_i,t} w_{a_j,t}}{\sqrt{\sum_{t\in T} w^2_{p_i,t}}\sqrt{\sum_{t\in T} w^2_{a_j,t}}}$$

The similarity of each web page and ad pair will be calculated. Then, for each page, ad messages will be sorted in order of its similarity to the targeted page. The ultimate ranking function will also take into account the keyword bid information. For each ad message that has a high bid (high CPC) would be ranked in priority. The ranking function will have to balance between the relevance and the keyword bid information. In general, an advertising system would try to gain the best effective cost per mille (1,000 impressions), which is calculated as:

Effective CPM = CPC * CTR

Where CPC is the keyword bid information and CTR is often associated with the relevance of the content of the targeted web page and an ad message.

3.5. Main Advantages of the framework

We have presented a general framework for contextual advertising that can produce a high quality match. Below we shall further detail or sum up the main advantages of this framework.

First, the framework is easy to implement and can efficiently rank ad messages based on their relevance to the targeted web page. In order to build a real-world content-targeted advertising system, we only have to choose and collect a large dataset called universal dataset, which is available and not "expensive" to get on the internet. The universal dataset should be general enough to cover all topics that would be mentioned in both web pages and ad messages.

Second, it can overcome one of the biggest problems in contextual advertising, the difference between vocabularies of web pages and ads. As discussed by Ribeiro et al [13], ad messages are often short, concise and general, whereas web pages can be about any topics with many specific terms. Moreover, a good advertisement for a web page is sometimes about a topic that is not mentioned explicitly in the web page. By analyzing topics for both web pages and ad messages, we can expand their vocabularies with the topics and hence, improve the relevance of a page and an ad that share the same topics.

Therefore, the framework can suggest appropriate ad messages for a targeted web page that have the same topics, thus share the same target audience.

Another important issue of the framework is that it can capture the semantic relations behind the content of web pages and ad messages. We have experienced the miss-match because of the homonym or multiple meaning of words while matching using tf.idf feature only. For example, a web page about cosmetic and skin cream (dưỡng da) was matched with an advertisement about leather shoes (da giày) because of the lexical misunderstanding. They are totally different but were matched because of the multiple-meaning word "da". Our system can mainly avoid this miss-match by taking into account the semantic factor that prioritizing ad messages which are topically related to the web page. In other words, it can reduce uncommon words and make the data more topic-focused.

Many studies recently have attempted to exploit the external large data that is available to use throughout the internet, such as semi-supervised learning. Our framework also takes advantage of such external large data in order to determine the semantic relatedness of words and documents in a wide domain.

Finally, our framework is flexible and general enough to be applied in different domains and different languages.

3.6. Chapter Summary

In this chapter, we have presented a general framework for contextual advertising with the support of the analysis of a large scale dataset. The main purpose is to improve the matching quality to suggest better advertisements for users based on their interest.

First, we prepare a large collection of data called Universal Dataset that can cover large enough topics and domains. We then use a hidden topic model to analyze it. After the estimation process, we use this model to do topic inference for web pages and ads. Eventually, pages and ads are matched after being enriched with hidden topics using cosine similarity.

Our framework can produce a high quality matching function for contextual advertising. It can reduce the miss-match by analyzing topics for web pages and ads. It overcomes one of the most difficult problems in contextual advertising: the difference

between vocabularies of web pages and ads (ads are often short and concise while web pages are in a bigger scope). The framework is also easy to implement, general and flexible enough to be applied in a multilingual environment for a real world contextual advertising system.

Chapter 4: Hidden Topic Analysis of Large-scale Vietnamese Document Collections

This chapter brings in-detail description of hidden topic analysis of large scale Vietnamese dataset [25] in the framework described in chapter 3. Section 4.1 presents hidden topic analysis, its background knowledge and theory. We then focus on Latent Dirichlet Allocation (LDA), a well-known hidden topic model that we choose to use in this application. Section 4.3 will describe in-detail our work on hidden topic analysis of a Vietnamese e-newspapers dataset, VnExpress data collection [8].

4.1. Hidden Topic Analysis

4.1.1. Background

Representing text corpora effectively in order to exploit their inherent essential relationship between members of the collections have become sophisticated over the years. There have been many studies aiming at modeling text documents recently.

One of the earliest methods proposed by Salton and McGill in 1983 is Vector Space Model [36]. It has been widely used in information filtering, information retrieval, ranking and indexing. One of its applications is using the assumptions of document similarities theory to calculate the relevance between a keyword search and a document. By representing each document as a vector with a separate term corresponding to a dimension, we calculate the cosine of the angle between those vectors to decide their similarities. Our simple matching method described in chapter 3 basically depends on this approach.

Since its appearance, there have been some other models developed based on the Vector Space Model, such names as generalized vector space model, topic-based vector space model and latent semantic analysis (LSA) [35]. LSA was first introduced in 1988 by Scott Deerwester et al. It is also sometimes called latent semantic indexing (LSI) in the context of its application to information retrieval [22]. LSA approach uses a term-document matrix to discover the relations between terms and documents by means of some concepts. It is a statistical method that can discover the semantic information through representing words and documents as points in Euclidean space. In this chapter, we focus on an approach that has some similar aspects to LSA, but instead of representing

words and documents as points in space, it displays the semantic relations of words and documents in terms of probabilistic topics. Such kind of model is called topic model, which will be introduced in the below section.

4.1.2. Topic Analysis Models

Topic models [15] are a significant step forward in modeling text corpora. They are based upon the idea that each document is a probability distribution over topics and each topic, in turns, is a mixture distribution over words (figure 11). Representing words and documents as probability distribution has some important advantages in compared with a simple space model. It can provide a probability distribution over words that can pick out correlated terms (figure 11).

The underlying idea of topic models is a probabilistic procedure of generating new documents. First, to make a new document (1), we can choose a topic distribution for the document, that means the document is composed of different topics with different distribution (2). Then, in order to generate words for the document, we can choose some words randomly based upon the distribution of words over those chosen topics (3). The generative process of making new document is illustrated in figure 11.

Conversely, given a set of documents, we can discover a set of topics that are responsible for generating a document, and the distribution of words that belong to a topic. Statistical method has been applied to model the generative process to estimate some parameters.

Two example of topic analysis using latent models are Probabilistic latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA).

PLSA, also known as probabilistic latent semantic indexing (pLSI), is a statistical technique for the analysis of two-mode and co-occurrence data [33]. It was developed based on LSA, adding a probabilistic model. It models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions. However, as argued by Blei, Ng et al (2003) [15], although pLSI is a useful step toward probabilistic modeling of text, it is incomplete in that it is not well-defined probabilistic model at the level of documents. As a result, it leads to the problem in assigning the probability to a document outside of the training test. Moreover, it can lead to the linearly growth of number of parameters along with the size of the corpus.

LDA, on the other hand, is a more completed topic analysis model that can overcome those disadvantages. It is the topic model that we have employed to develop our contextual advertising framework. The detail of LDA is described in the following section.

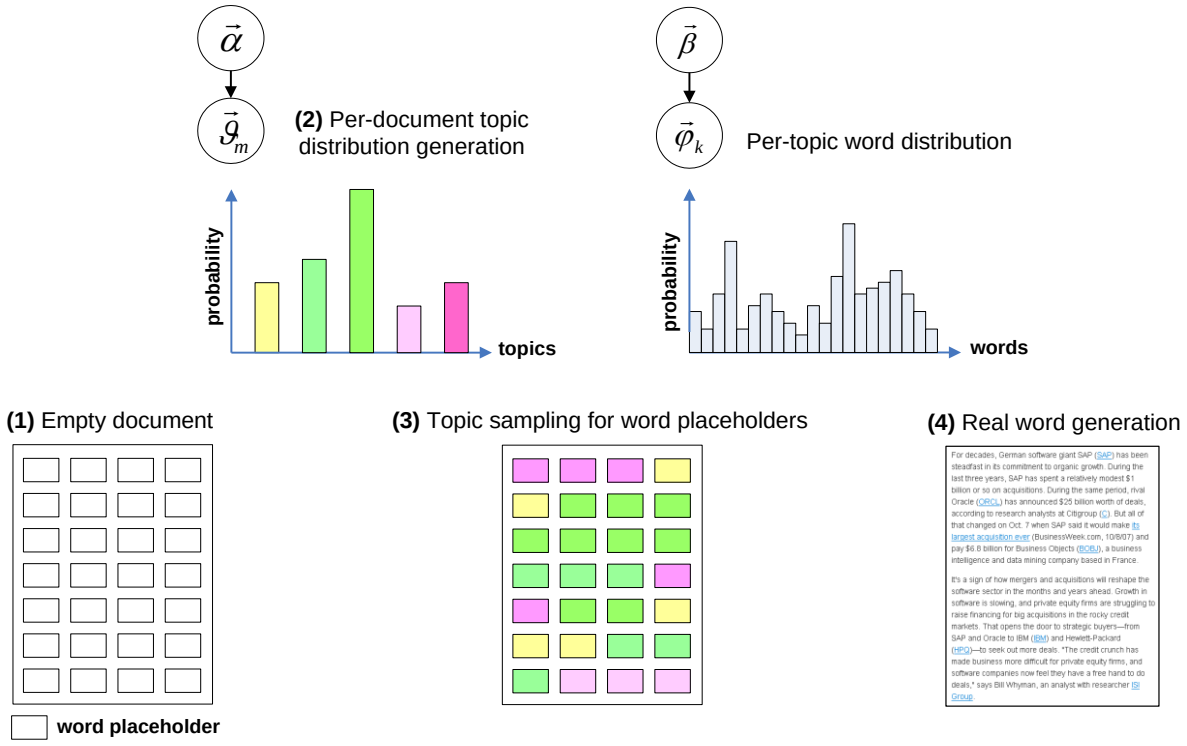


Figure 11: Generating a new document by choosing its topic distribution and topic-word distribution [32]

4.1.3. Latent Dirichlet Allocation (LDA)

LDA is a generative model introduced by Blei, Ng et al (2003) [15]. Similar to pLSA in the fundamental idea, LDA is also based upon the consideration that a document is a mixture of topics. However, it makes different statistical assumptions against pLSA.

It is a three-level Bayesian model (corpus level, document level and word level) (figure 12).

- **Generative process**:

Given a set of M documents: $D = \{d_1, d_2, ..., d_M\}$, the document m is composed of N_m words w_i with $w_i \in \{t_1, \ldots t_v\}$, where v is the total number of terms.

- α and β is the corpus-level parameters
- $\vec{\vartheta}_m$ is the topic distribution over document m (document-level parameter)
- $Z_{m,n}$ is the topic index of word n of document m (word-level variable)
- $\vec{\varphi}_{z_{m,n}}$ is the topic-specific term distribution of $z_{m,n}$

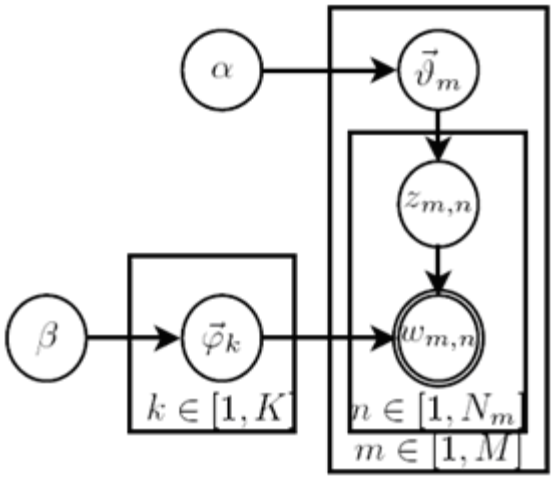


Figure 12. Graphical model representation of LDA - The boxes is "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Arrows indicate conditional dependencies between variables and white plates refers to repetitions of sampling steps (figure 12). The generative process is described as below:

⊡ "topic plate"
for all topics $k \in [1, K]$ **do**
  sample mixture components $\vec{\varphi}_k \sim \text{Dir}(\vec{\beta})$
end for
⊡ "document plate":
for all documents $m \in [1, M]$ **do**
  sample mixture proportion $\vec{\vartheta}_m \sim \text{Dir}(\vec{\alpha})$
  sample document length $N_m \sim \text{Poiss}(\xi)$
  ⊡ "word plate":
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n} \sim \text{Mult}(\vec{\vartheta}_m)$
    sample term for word $w_{m,n} \sim \text{Mult}(\vec{\varphi}_{z_{m,n}})$
  end for
end for

Since $w_{m,n}$ is conditional dependent on the distribution $\vec{\varphi}_k$ and $z_{m,m}$ is dependent on the distribution $\vec{\vartheta}_m$, we have the probability that a topic index $w_{m,n}$ is a word t belongs to the topic distribution over the document ($\vec{\vartheta}_m$) and the topic-term distribution ($\underline{\Phi}$):

$$p(w_{m,n} = t \mid \vec{\vartheta}_m, \underline{\Phi}) = \sum p(w_{m,n} = t \mid \vec{\varphi}_k)\, p(z_{m,n} = k \mid \vec{\vartheta}_m)$$

For the probability of a term, we can then calculate the probability of all the documents based on the Bayesian probability:

$$p(\vec{d}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi} \mid \vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n} \mid \varphi_{z_{m,n}})\, p(z_{m,n} \mid \vec{\vartheta}_m)}_{\text{wordplate}}.p(\vec{\vartheta}_m, \vec{\alpha})}^{\text{documentplate(1document)}}.\underbrace{p(\underline{\Phi} \mid \vec{\beta})}_{\text{topic plate}}$$

It consists of the topic plate, word plate and document plate. Once having the distribution $p(\vec{d}_m \mid \vec{\alpha}, \vec{\beta})$, we then can have the probability of the whole corpus $W = \{\vec{d}_m\}_{m=1}^{M}$ as the multiplication of all the above probabilities:

$$p(W \mid \vec{\alpha}, \vec{\beta}) = \prod_{m=1}^{M} p(\vec{d}_m \mid \vec{\alpha}, \vec{\beta}).$$

- **Parameter Estimation and Inference using Gibbs sampling**

In the generative model, we have discussed the way a new document generated. Conversely, given a set of documents, the purpose of this process is to discover the topic model that has generated all these documents. This task includes: first, estimating the topic-word distribution $\vec{\varphi}_k$; second, estimating the topic distribution $\vec{\vartheta}_m$ for each document. Given the observed words w, our task is to estimate the model and do the inference for a new document by that model. We will describe an algorithm that using Gibbs sampling, a form of Markov chain Monte Carlo to deal with that.

The process of estimating parameters is composed of two steps: initialization and burn-in period.

- **Initialization**

zero all count variables, $n_m^{(z)}$, n_m, $n_z^{(t)}$, n_z

for all documents $m \in [1, M]$ do

for all words $n \in [1, N_m]$ in document m do

sample topic index $z_{m,n}$ ~Mult(1/$\boldsymbol{K}$)

increment document-topic count: $n_m^{(s)} + 1$

increment document-topic sum: $n_m + 1$

increment topic-term count: $n_s^{(t)} + 1$

increment topic-term sum: $n_z + 1$

end for

end for

increment topic-term sum: $n_z + 1$

end for

end for

- **Burn-in period**

while not finished do

for all documents $m \in [1, M]$ **do**

for all words $n \in [1, N_m]$ in document m **do**

- for the current assignment of z to a term t for word $w_{m,n}$:

decrement counts and sums: $n_m^{(z)} - 1$; $n_m - 1$; $n_z^{(t)} - 1$; $n_z - 1$

- multinomial sampling acc. (decrements from previous step):

sample topic index $\tilde{z} \sim p(z_i \mid \vec{z}_{-i}, \vec{w})$

- use the new assignment of z to the term t for word $w_{m,n}$ to:

increment counts and sums: $n_m^{(\tilde{z})} + 1$; $n_{\tilde{z}}^{t} + 1$; $n_{\tilde{z}} + 1$

end for

end for

Where $n_m^{(z)}$: the number of topic z in document m

n_m: the total number of topics in document m

$n_z^{(t)}$: the number of term t in topic z

n_z : the total number of terms in topic z

In burn-in period, every parameter is sampled again till it reaches a reasonable precision. For each sampling iteration, the parameters corresponding to each term and topic are adjusted. When the period is finished, parameters are read out. Two hidden distributions are calculated as:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v} \qquad \vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{z=1}^{K} n_m^{(z)} + \alpha_z}$$

After estimating the model, we then can use this model to do topic inference for new documents.

4.2. Process of Hidden Topic Analysis of Large-scale Vietnamese Datasets

4.2.1. Data Preparation

With the purpose of using a large scale dataset for Vietnamese contextual advertising, we choose VnExpress [8] as the dataset for topic analysis. VnExpress is one of the highest ranking e-newspaper in Vietnam, thus contains a large number of articles in many topics in daily life. For this reason, it is a suitable dataset for advertising areas.

The dataset includes different topics, such as Society, International news, Lifestyle, Culture, Sports, Science, etc. We crawled 220 Megabyte of approximately 40,000 pages using Nutch [28]. After preparing the data, we then do preprocessing for them.

4.2.2. Data Preprocessing

Data preprocessing plays an important role in improving the performance of the overall framework. This task includes the following steps (figure 13):

- HTML remover
- Sentence Segmentation: To separate different sentences. Normally, each sentence is separated by full stop, question mark or exclamation mark (.?!). However, we have to discriminate between the signal of sentence separation and other cases (such as full stop in: decimal numbers, email address, name title such as Mr., Ms., Dr., etc.)

- Sentence Tokenization: To detach marks from their previous words (,) or sentences (.?!).
- Word Segmentation: Vietnamese is often considered as monosyllabic; hence a word might be composed of more than one syllable. This leads to the problem of word segmentation. In this process, we combine two or more syllables in a word as one to separate different words by white space.
- Filters: removing tokens after word segmentation, such tokens as number, date/time, and too short tokens. Too short sentences or Vietnamese sentences without tones should also be removed.
- Remove trivial or non topic-oriented words: such as functional words, too rare or too common words.

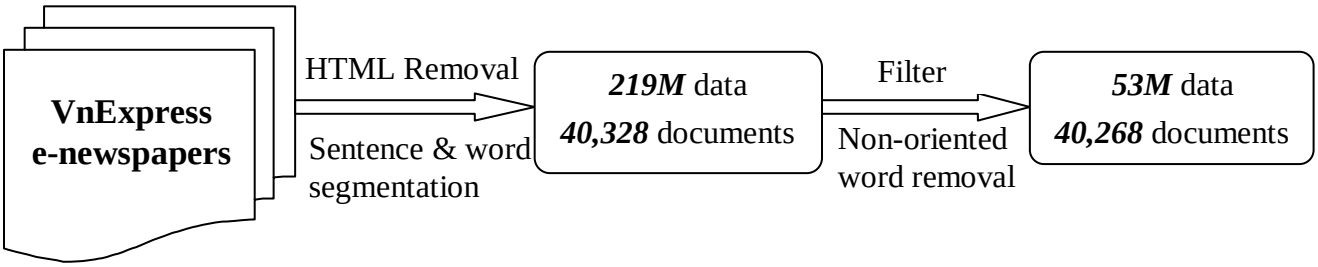


Figure 13: VnExpress Dataset Statistic

The above mentioned tasks were done using the toolkit JVnTextPro [25] for preprocessing.

4.3. Hidden Topic Analysis of VnExpress Collection

After preprocessing, we analyze hidden topics for the dataset using JGibbsLDA [26] [31], a toolkit of LDA and Gibbs Sampling. We examine the results of hidden topic analysis of three models: 60, 120 and 200-topic models respectively. Here are some samples of hidden topics analyzed through the dataset in 200-topic model[7] (table 2).

[7] For the full results of hidden topic analysis of 60,120, 200-topic models, see online at:
http://gibbslda.sourceforge.net/vnexpress-060topics.txt
http://gibbslda.sourceforge.net/vnexpress-120topics.txt
http://gibbslda.sourceforge.net/vnexpress-200topics.txt

200-topic model

Topic 1	Topic 3	Topic 15	Topic 44
phòng (room)	bác_sĩ (doctor)	thời_trang (fashion)	thiết_bị (equipment)
không_gian (space)	bệnh_viện (hospital)	người_mẫu (model)	sản_phẩm (product)
thiết_kế (design)	thuốc (medicine)	mặc (wear)	máy (machine)
ngôi_nhà (house)	bệnh (disease)	trang_phục (clothes)	màn_hình (screen)
tầng (floor)	phẫu_thuật (surgery)	thiết_kế (design)	công_nghệ (technology)
trang_trí (decorate)	điều_trị (treatment)	đẹp (beautiful)	điện_thoại (telephone)
nội_thất (interior)	bệnh_nhân (patient)	váy (dress)	hãng (company)
tường (wall)	y_tế (medical)	sưu_tập (collection)	sử_dụng (use)
ánh_sáng (light)	ung_thư (cancer)	mang (wear)	thị_trường (market)
đèn (lamp)	tình_trạng (condition)	phong_cách (style)	Usd (USD)
phòng_ngủ (bedroom)	cơ_thể (body)	quần_áo (costume)	pin (battery)
rộng (wide)	sức_khoẻ (health)	nổi_tiếng (famous)	cho_phép (allow)
bố_trí (arrange)	đau (hurt)	quần (trousers)	samsung (samsung)
vườn (garden)	gây (cause)	trình_diễn (perform)	di_động (mobile)
kính (glass)	khám (examine)	thích (like)	sony (sony)
cảm_giác (feel)	kết_quả (result)	quyến_rũ (charming)	nhạc (music)
diện_tích (square)	căn_bệnh (illness)	sang_trọng (luxurious)	máy_tính (computer)
căn_phòng (apartment)	nặng (serious)	vẻ_đẹp (beauty)	hỗ_trợ (support)
khu (area)	cho_biết (inform)	gái (girl)	điện_tử (electronic)
hiện_đại (modern)	máu (blood)	gương_mặt (figure)	tính_năng (feature)
cầu_thang (stair)	xét_nghiệm (test)	siêu (super)	kết_nối (connect)
phòng_khách (living-room)	chữa (cure)	áo_dài (aodai)	thiết_kế (design)
căn_hộ (flat)	chứng (trouble)	giày (shoes)	chức_năng (function)

Topic 48	Topic 56	Topic 172	Topic 45
chứng_khoán (stock)	bánh (cake)	thẻ (card)	gồm (comprise)
công_ty (company)	mcdonald (McDonald)	khoá (lock)	logo (logo)
đầu_tư (invest)	thịt (meat)	rút (withdraw)	mang (hold)
ngân_hàng (bank)	pizza (pizza)	chủ (owner)	hình_ảnh (image)
cổ_phần (joint-stock)	ba_tê (pate)	chìa (key)	tre (bamboo)
thị_trường (market)	bánh_mì (bread)	thẻ_tín_dụng (credit card)	mây (rattan)
giao_dịch (transaction)	bánh_ngọt (pie)	Atm (ATM)	biểu_tượng (symbol)
đồng (dong)	cửa_hàng (shop)	tín_dụng (credit)	thể_hiện (show)
mua (buy)	xúc_xích (hot dog)	thanh_toán (pay)	xu_hướng (trend)
phát_hành (publish)	kem (ice-cream)	visa (visa)	thủ_công (handicaft)

niêm_yết (post)	khai trương (open)	tối_thiểu (minimum)	trang_trí (decorate)
bán (sell)	nguội (cold)	mastercard	truyền_thống (traditional)
tài_chính (finance)	hamburger (hamburger)	phát_hành (release)	rối (puppet)
đấu_giá (auction)	mcdonald (McDonald)	trả_nợ (pay_debt)	bóng (ball)
trung_tâm (center)	thịt (meat)	sẵn_sàng (ready)	bàn_tay (hand)
thông_tin (information)	nhà_hàng (restaurant)	mật_mã (password)	nhân_vật (character)
doanh_nghiệp (business)	đồ_ăn (food)	thường_niên (annual)	sáp (wax)
cổ_đông (shareholder)	sandwich (sandwich)	cảnh_giác (alert)	nghệ_nhân (artisan)
nhà_đầu_tư (investor)	khẩu_vị (taste)	chủ_thẻ (owner)	phong_cách (style)
nhà_nước (government)	tiệm_bánh (bakery)	theo_dõi (follow)	thiết_kế (design)
tổ_chức (organization)	bảo_đảm (ensure)	nhà_băng (bank)	gối (pillow)
triệu (million)	nướng (grill)	tội_phạm (criminal)	vòng (round)
quỹ (budget)	bí_quyết (secret)	trộm (steal)	tạo_nên (make)

Table 2: An illustration of some topics extracted from hidden topic analysis

Table 2 shows some sample topics derived from 200-topic model, the topics are typically as interpretable as the ones shown here. Each word has a distribution over the corresponding topic. It has shown a satisfactory result and can be extremely useful in many applications.

4.4. Chapter Summary

In this chapter, we have introduced about topic models with focus on LDA, a well-known topic model that its potential has been proved through many applications [32]. We then apply this method for topic analysis of a large scale Vietnamese dataset, VnExpress. We examine the analysis of this dataset with different number of topics: 60, 120 and 200, which have shown satisfactory results. The deployment of these topic models and their potential contribution in our framework will then be presented in the next chapter.

Chapter 5: Evaluation and Discussion

In contextual advertising, matching and ranking ad messages based on their relevance to the targeted web page are an important factor. As stated earlier, it increases the likelihood of visits to the website pointed by the ad. In chapter 3, we have introduced our framework to perform this task. In order to evaluate the performance of the framework, we carry out different experiments that will be presented in-detail below.

5.1. Experimental Data

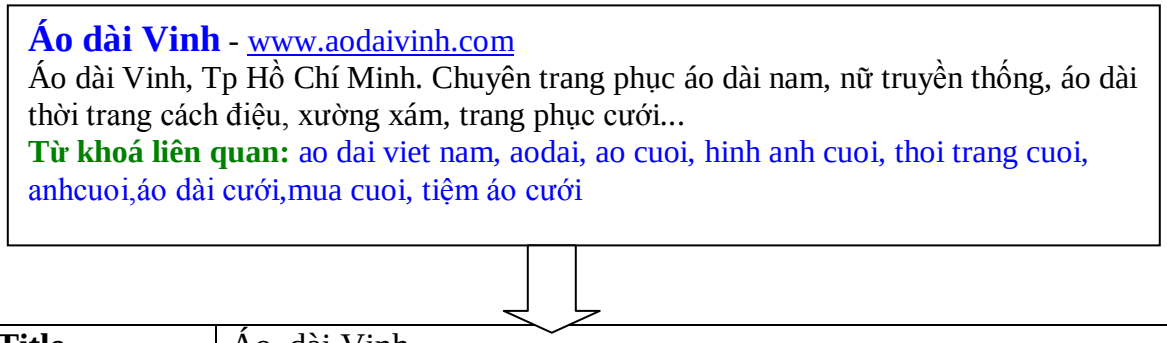
We quantify the effect of matching using hidden topics and without hidden topics using a set 100 web pages and 2,607 unique ad messages (figure 15).

- For web pages, we choose 100 pages randomly from a set of 27,763 pages crawled from VnExpress e-newspapers, one of the highest ranking Vietnamese e-newspapers [8]. Those pages are chosen from different topics: Food, Shopping, Cosmetics, Mom & children, Estate, Stock, Jobs, Law, etc. These topics are primarily classified on the e-newspaper. And note that the information of these classified topics is not used in our experiments, just for reference here only.

- For advertising messages, as contextual advertising has not yet been applied in Vietnam to our knowledge, it is difficult to find a real Vietnamese advertisement collection. As stated in chapter 2, advertisement's type in Vietnam is mainly banners, thus such kind of real ad messages might be not available. We have also contacted some online advertising companies, such as VietAd [5] , a company which keyword-based advertising system had once been tested in VietnamNet [10]. However, their database was just for tested and the number of such advertisements was only a few (less than 10 ad messages).

In order to perform the experiments, we choose another resource: Vietnamese websites directory [9] . It suits the form of ad messages perfectly. Real advertisement database of some contextual advertising systems, such as Google and Yahoo, also has the similar form. We assume that each website, its title, description and keywords are an ad message. Therefore, an ad message is composed of four parts: title, website's URL, its description and keywords. Figure 14 is an example of an ad message in our experimental data.

After crawling all 3,982 ad messages, we did the preprocessing and transformation including sentence segmentation, sentence tokenization, filters and non topic-oriented words removal. We used JVnTextPro toolkit [25] for this work. It is similar to the work described in section 4.2.2.



Title	Áo_dài Vinh
URL	www.aodaivinh.com
Description	Áo_dài Vinh , Tp_Hồ_Chí_Minh. Chuyên trang_phục áo_dài nam , nữ truyền_thống , áo_dài thời_trang cách_điệu , xường_xám , trang_phục cưới...
Keywords	áo_dài việt_nam , áo_dài , áo_cưới , hình_ảnh cưới , thời_trang cưới , ảnh_cưới , áo_dài cưới , mùa_cưới , tiệm áo_cưới

Figure 14: An advertisement message, before and after preprocessing

Nevertheless, keywords in this database are almost none-tone, so we cannot use them to enhance the matching performance. However, keywords play an important role in contextual advertising. The contribution of them in matching task has been proved through experiments and affirmed in many works [13] [12] [23]. Therefore, we decide to recover tone for all keywords of each ad messages in order to improve the contextual matching.

- **Tone recovery:** Since the number of ad messages is large and they contain a lot of Vietnamese none-tone keywords, it is a time-consuming task to recover tone for all of them manually. In order to make it easier, we first list all of the distinct keywords, which are over 2,500 in all the set of ad messages since there are many overlap keywords. We then correct and recover tones for this set.

To apply these corrected keywords for each ad message, we match its each keyword to the corrected keyword – that has been tone-removed. If they are matched, we then change that keyword to the corresponding corrected one.

By doing so, we have already corrected all keywords for each ad messages.

After recovering tone, selecting and filtering the data, there are 2,607 ad messages left. Each ad message has the same form as illustrated in figure 14. The domains of ad messages are various, such as Educations, Music, Films, Economics, Government, Computers, etc. We then use all of them for matching with 100 selected pages above (figure 15).

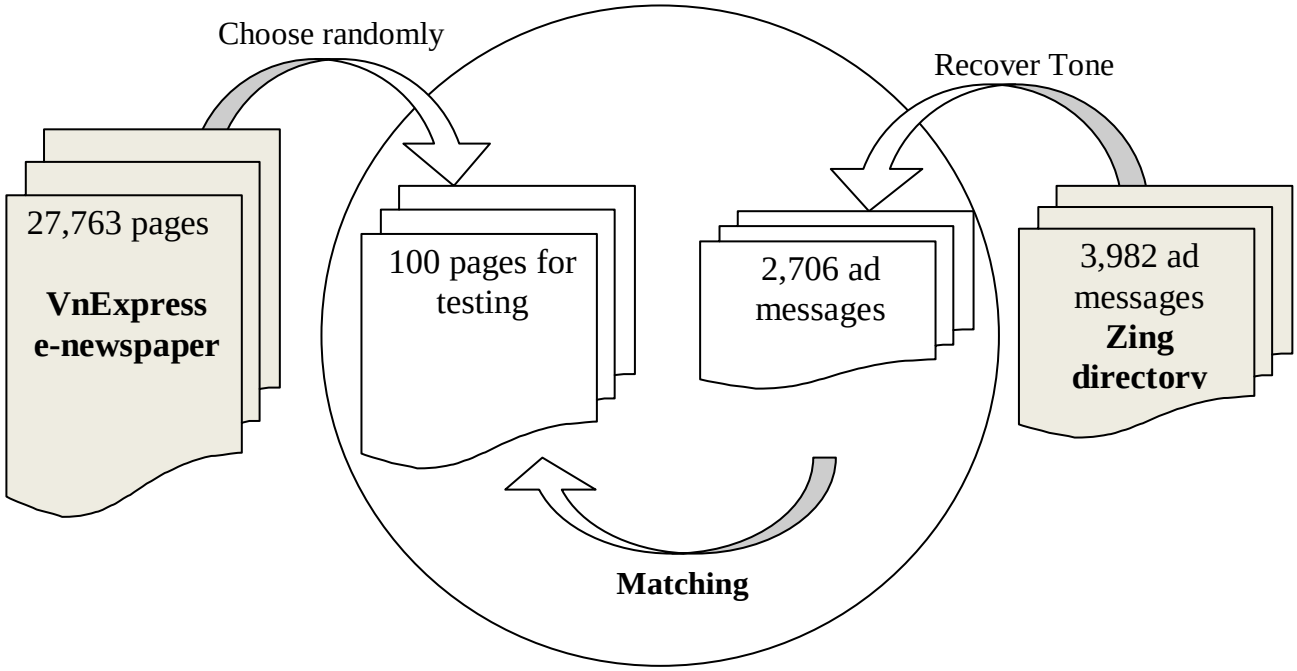


Figure 15: Webpage and Advertisement Dataset Statistic

5.2. Parameter Settings and Evaluation Metrics

In order to evaluate the importance of keywords in contextual matching and the contribution of hidden topics in this framework, we perform some different matching strategies as follows:

First, to assess the impact of keywords in contextual matching, we implement two retrieval baselines following the approach of Ribeiro-Neto et al, 2005 [13]. The first

strategy is called AD, that means matching the web page and ad message using its title and description only. The second one is AD_KW, matching the web page and ad message also using its keywords, which have already been tone-recovered. Then, the similarity of a webpage p and an ad message a is defined as:

$sim(AD) = sim(p, a)$

$sim(AD\_KW) = sim(p, a \cup kw)$

where *kw* is the keywords belonging to the ad message a.

We then use the winning strategies as the baseline for our later method using hidden topics.

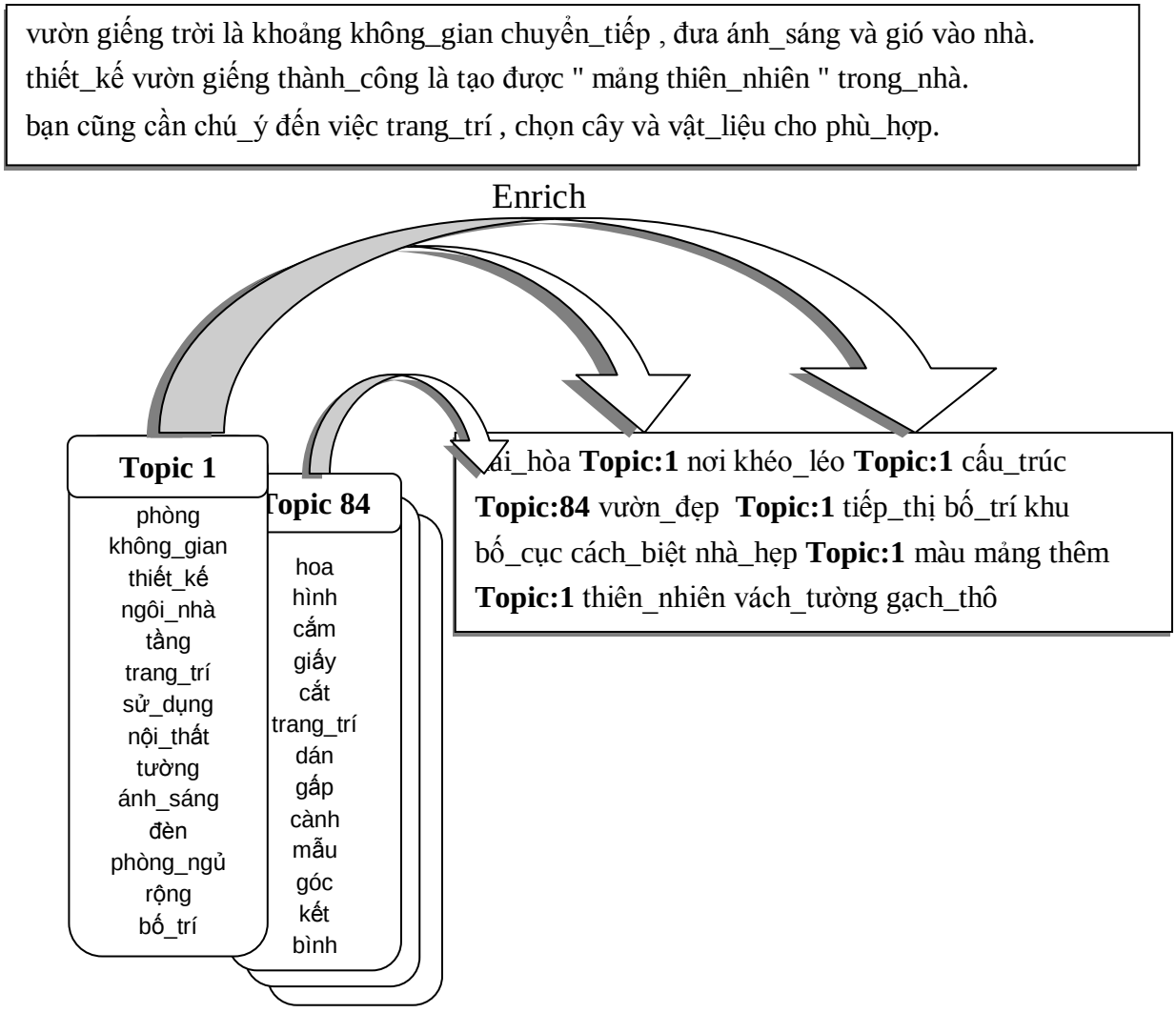


Figure 16: Example of an ad before and after being enriched with hidden topics - Some most likely words in the same hidden topics.

Second, to evaluate the contribution of hidden topics, we carry out six different experiments, which are called HT strategies. After doing topic inference for all web pages and ads, we expand their vocabularies with their most likely hidden topics.

As described in chapter 3, each web page or ad will have a distribution over topics. We then choose topics that have high distribution to enrich that page or ad. Example of an ad message that has been enriched with hidden topics is illustrated in figure 16.

In these experiments, we use estimated models with 60, 120 and 200 topics in turns. The hidden topic analysis of the dataset to these estimated models has been discussed detail in section 4.3. We also use two different levels of expanding with hidden topics to decide the number of topics added to a webpage/ad. In particular, we use two parameters: cutoff and scale as follows:

$$\text{NumberTopic}_d(t) = \begin{cases} \text{round}(\text{scale}.\text{dis}_d(t)) & \text{if } \text{dis}_d(t) >= \text{cutoff} \\ 0 & \text{if } \text{dis}_d(t) < \text{cutoff} \end{cases}$$

where $NumberTopic_d(t)$ is number of times topic t added to the document d.

$dis_d(t)$ is the distribution of topic t in the document d.

cutoff is the topic distribution threshold.

scale is the parameter that determines the number of hidden topics added.

In our experiments, we use the value *cutoff* = 0.05 and try two different scales: 10 and 20. For example, if $dis_d(t)$ = 0.1 and *scale* = 20, then topic t will be added twice to the document d.

Our six matching experiments using hidden topics will be called HTx_y for short from now on, where x stands for the number of hidden topics of the used estimated model and y is the scale. We therefore perform six experiments:

HT60_10, HT60_20, HT120_10, HT120_20, HT200_10 and HT200_20 (Table 3)

Methods	Descriptions
AD	Matching using title and description of ads
AD_KW	Matching using title, description and keywords of ads
HT60_10	Matching using hidden topics – number of topics: 60, scale: 10
HT60_20	Matching using hidden topics – number of topics: 60, scale: 20
HT120_10	Matching using hidden topics – number of topics: 120, scale: 10
HT120_20	Matching using hidden topics – number of topics: 120, scale: 20
HT200_10	Matching using hidden topics – number of topics: 200, scale: 10
HT200_20	Matching using hidden topics – number of topics: 200, scale: 20

Table 3: Description of 8 experiments without hidden topics (AD and AD_KW) and with hidden topics (HT)

- To evaluate the performance of the matching method using retrieval information (term frequencies) only and the matching method using hidden topics, we prepare the test data as follows:

First, we start by matching each webpage to all the ad messages and rank them to their similarities. Each method, AD, AD_KW, HT60_10, HT60_20, HT120_10, HT120_20, HT200_10 and HT200_20, will propose a different rank list of ad messages to a targeted page. Since the number of ad messages is large, these lists can be different from this method to another method with little or no overlap.

In order to determine the precision of each method and compare them, we first select top four ranked ads of each method and put them to a pool for each targeted page. Consequently, each pool will have no more than 32 ad messages. We then select from these pools most relevant ads and exclude irrelevant ones. In average, each web page will be matched with 6.51 ads eventually. The total number of web pages is 100 (figure 17).

This evaluation method is similar to the literature of Ribeiro-Neto et al, 2005 [13], Lacerda et al, 2006 [12], and Ciaramita et al, 2008 [23].

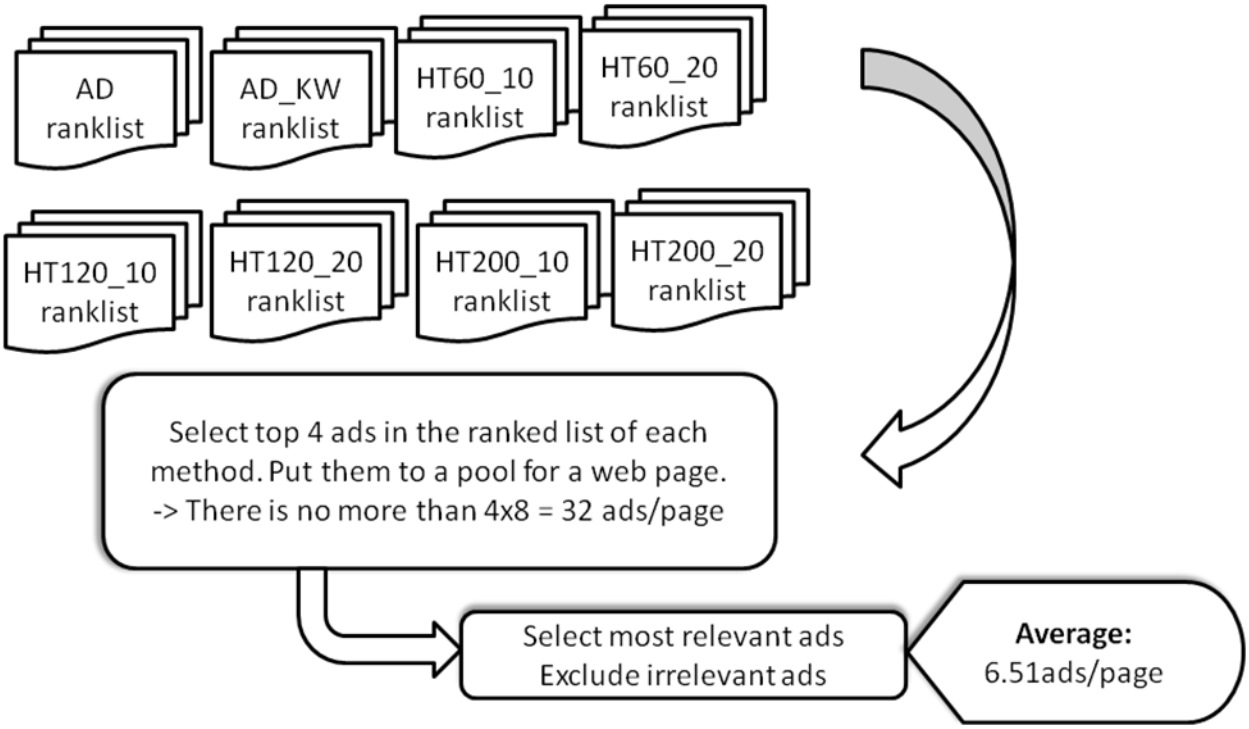


Figure 17: Selecting top 4 ads in each ranked list for each corresponding webpage for evaluation

To calculate the precision of each method, we use 11 point average score, a performance measurement that is often used for a ranking system. The algorithm to calculate this 11 point average value was introduced by Yang, 1999 [34]. The detailed algorithm will be described as follows:

- **11 point average precision:**

For every ad message rank lists, we calculate the precision at every 11 point of recall: 0, 0.1, 0.2,..., 0.8, 0.9, 1.0. Finally, the average precision of these 11 points is returned [39]. The recall and precision in this system are calculated as:

$$precision = \frac{\text{ad messages found and correct}}{\text{total ad messages found}}$$

$$recall = \frac{\text{ad messages found and correct}}{\text{total ad messages correct}}$$

Let pre(r) be the precision at recall = r. We need to calculate 11 values of pre(r) with

$r \in R$, $R = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. The algorithm is as follows:

1. For each webpage, calculate the precision and recall at each position in the ranked list where a correct ad message is found.

2. For each interval between thresholds 0, 0.1, 0.2,..., 0.8, 0.9, 1.0, use the highest precision value in that interval as the "representative" precision value at the left boundary of this interval.

$pre(r) = \max \{pre(i) \mid i \in [r, r + 0.1]\}$

For example: If there are two precision values in the interval [0.1, 0.2]:

$pre(0.15) = 0.4$ and $pre(0.17) = 0.5$,

then $pre(0.1) = \max\{pre(0.15), pre(0.17)\} = 0.5$.

3. For the recall threshold of 1.0 the "representative" precision is either the exact precision value if such point exists, or the precision value at the closest point in terms of recall. If the interval is empty we use the default precision value of 0.

4. ***Interpolation:*** At each of the above recall thresholds replace the "representative" precision using the highest score among the "representative" precision values at this threshold and the higher thresholds.

$pre(r) = \max \{pre(i) \mid i \in [r,1]\}$, therefore:

$\forall x, y \in R, (x \geq y) \Leftrightarrow pre(x) \geq pre(y)$.

5. ***Per-interval averaging:*** Average per-document data points over all the test documents at each of the above recall thresholds respectively. This step results in 11 per-interval precision scores.

6. ***Global averaging:*** Average of the per-interval average precision scores to obtain a single-numbered performance average. The resulting value is called the 11-point average precision.

For each method, we calculate the per-interval average and global average (11-point score) to compare them. Additionally, to quantify the ranking quality of each method, we

use the number of correct ad messages found at position 1, 2 and 3 (#1, 2, 3) in each ad message rank lists.

5.3. Experimental Results

First, we analyze the impact of keywords in contextual matching by calculating the 11 precisions of the corresponding recall. Our result is illustrated in figure 18. The AD_KW method performs better than the simple match strategy using title and description of ad messages only.

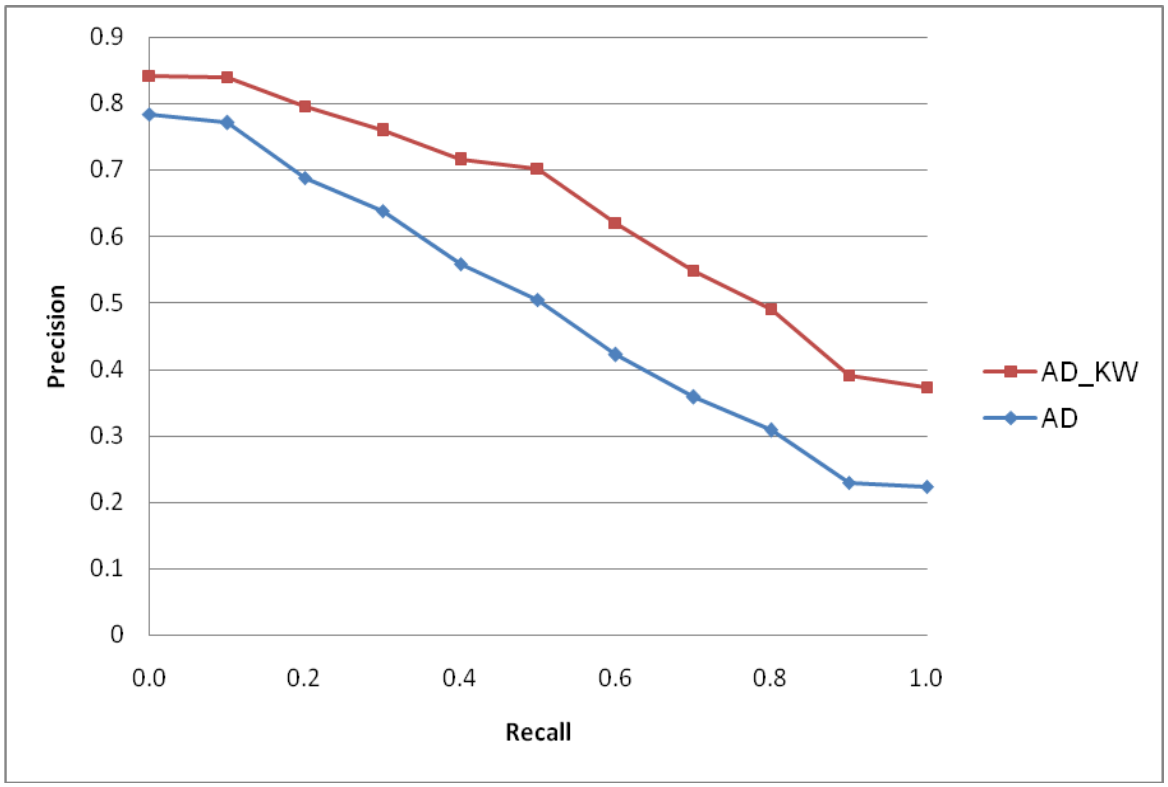


Figure 18: Precision and Recall of matching without keywords (AD) and with keywords (AD_KW)

We then use the better method (AD_KW) as the baseline for our next experiments using hidden topics. We examine the contribution of hidden topics using different estimated models: the model of 60, 120 and 200 topics. The result for those methods using hidden topics (HT) are displayed in figure 19, in compared with the precisions of the baseline method (AD_KW).

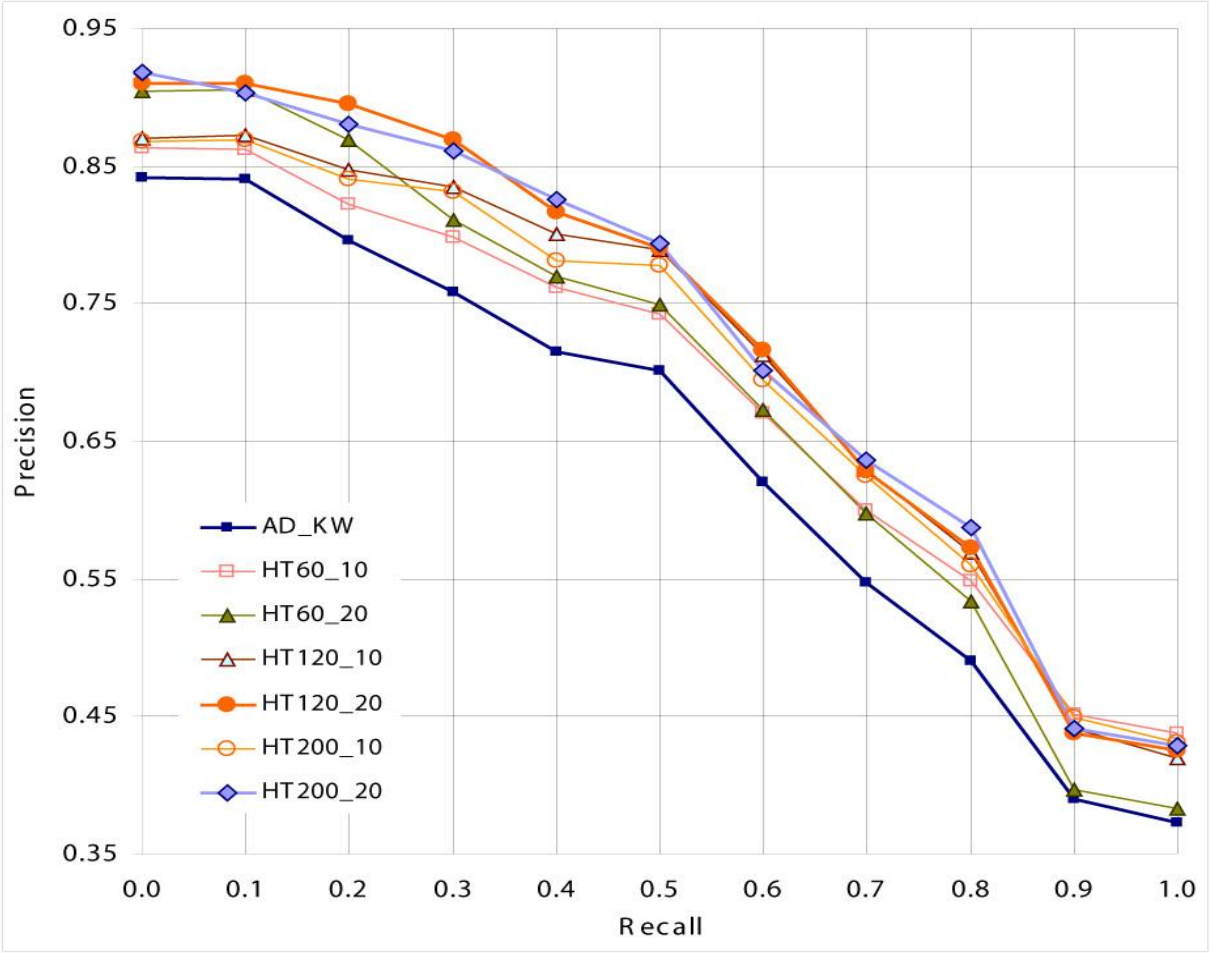


Figure 19: Precision and Recall of matching without hidden topics (AD_KW) and with hidden topics (HT)

We also calculate the number of corrected ad messages found in position 1, 2 and 3 (#1, 2, 3) in the rank list of each method. Table 4 shows the result and summarizes the performance of all 8 methods by their 11 points average precisions.

Contextual matching using hidden topics has shown a potential result. If using keywords has increased the precision considerably, it still has some critical miss-match in some cases. Example of such miss-match is illustrated in figure 20 and 21. Top 3 proposed ad messages for the targeted page of the method HT200_20 and the analysis of word co-occurrence and topic analysis of the example are displayed in figure 21.

Methods	Corrected Ads found				11-points average precision
	#1	#2	#3	Totals	
AD	70	56	52	178	**49.86%**
AD_KW	78	69	64	211	**64.32%**
HT60_10	79	76	70	225	**68.72%**
HT60_20	86	75	67	228	**69.02%**
HT120_10	82	74	74	230	**70.76%**
HT120_20	89	79	69	237	**72.47%**
HT200_10	79	77	77	233	**70.26%**
HT200_20	88	78	79	245	**72.50%**

Table 4: Precision at position 1, 2, 3 and the 11-points average score

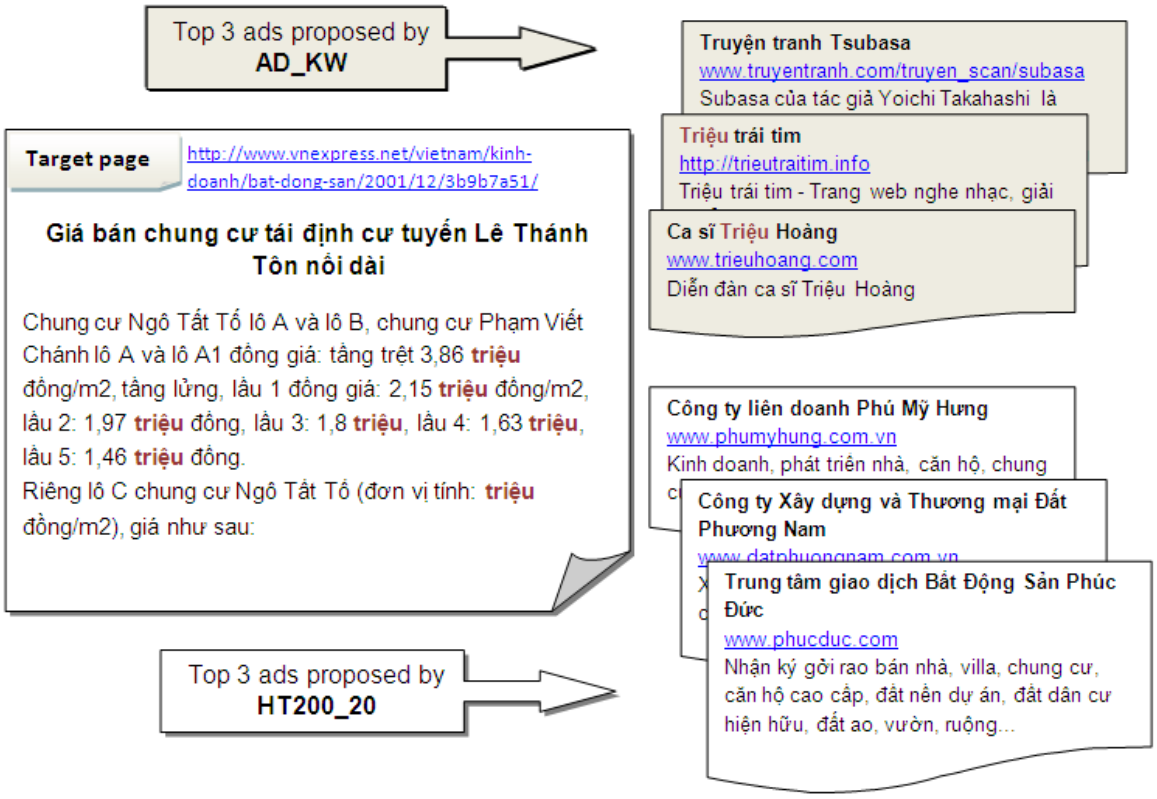


Figure 20: Sample of matching without hidden topics (AD_KW) and with hidden topics (HT200_20)

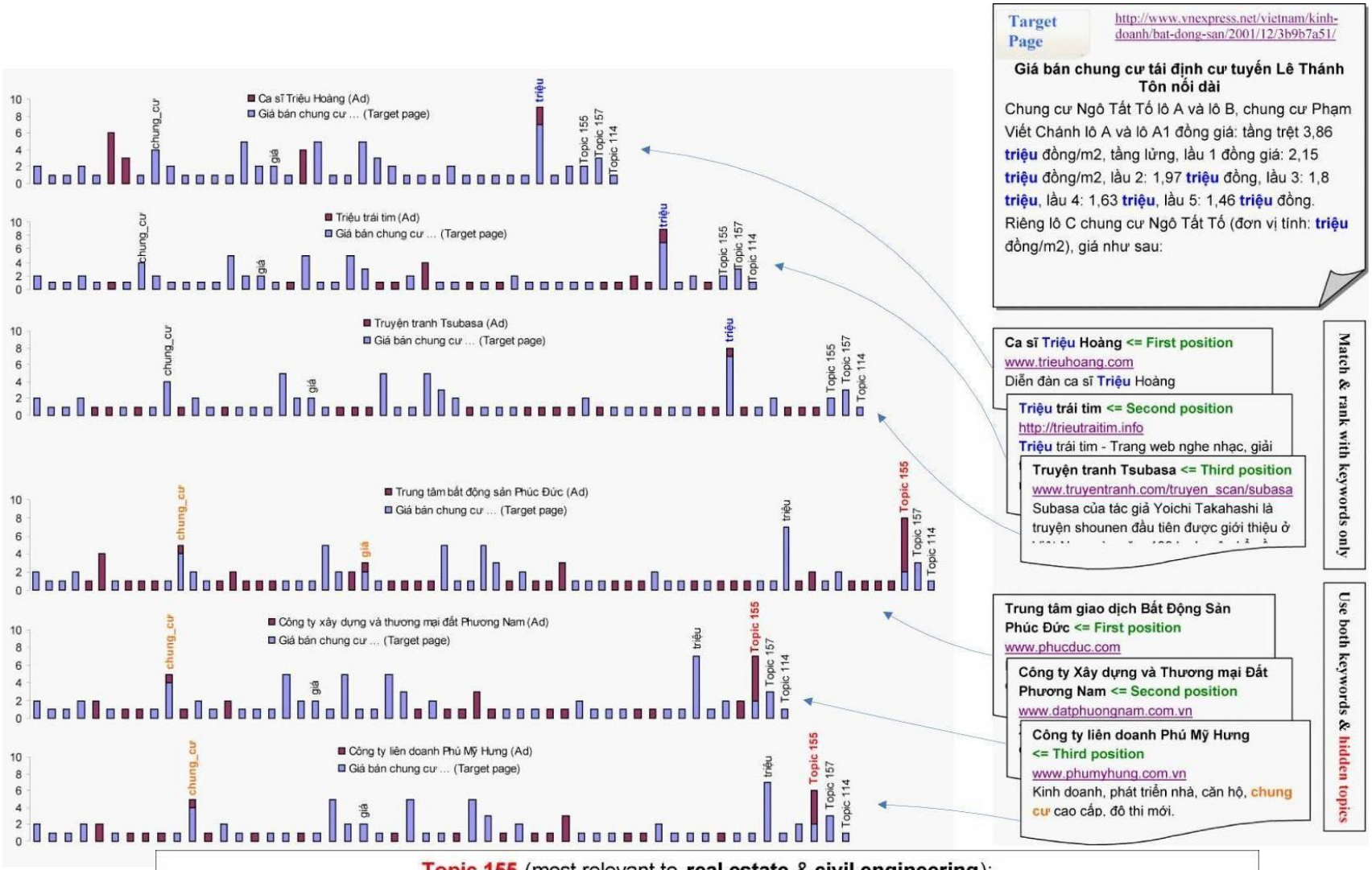


Figure 21: Word co-occurrence vs. Topic distribution of targeted page and top 3 ad messages proposed by HT200_20 in figure 20

5.4. Analysis and Discussion

In our work, we have examined the importance of keywords and its contribution in contextual matching. We then evaluate the performance of matching method using hidden topics as our key idea and approach to this problem.

The first experiment of simple matching using title and description of ad messages and then the contribution of keywords reflect the importance of keywords in this model. It can be explained that: keywords, which are important and specific words that have been chosen by advertisers, are often representatives of the whole page. It therefore gives a straightforward way for matching with the targeted page.

Following this idea, we realize that an expand of vocabularies in both web pages and ad messages can yield a better result. Moreover, semantic relations are also an importance factor in this contextual matching and ranking problem. We therefore examine the effect of hidden topics in this application with the better method carried out in the first experiment as the baseline. As shown in figure 19, using hidden topics has significantly improved the performance of the whole framework. It increases the precision in average from 64% to 72% and reduces almost 23% error.

In particular, we verify the contribution of topics in many cases that normal matching strategy cannot find appropriate ad messages for the targeted pages. Since retrieval matching is based on only the syntactic feature of web pages and ads, it is sometimes deviated by unimportant words that are not practical in matching. An example of such case is illustrated in figure 20. The word “triệu” (million) is repeated many times in the targeted page, hence given a high weight in syntactic matching. The system then misleads in proposing relevant ad messages for this targeted page. It puts ad messages having the same high-weighted word “triệu” in the top ranked list. However, in our method, it has shown a satisfactory result and can overcome such miss-match. As illustrated in figure 21, those three ads proposed by our system do not share many words with the targeted page. However, by analyzing topics for both of them, we can find out their latent semantic relations and thus realize their relevance. This has magnificently overcome the problem of miss-match in retrieval matching method.

For the overall methods, we also calculate the number of corrected ad messages found in the first, second and third position of the rank lists proposed by each strategy (#1,

#2, #3 in table 3). Since in contextual advertising, normally, we only consider some first ranked ads, we want to examine the precision of these top slots. It also reflects the precision of our hidden-topic methods higher than that of the baseline matching method. Moreover, the precision at position 1 (#1) is generally higher than that of position 2 and 3 (#2, #3). If the system is ranking the relevant ads near the top of the ranked list, it is possible that the system can suggest most appropriate ads for the corresponding page. It therefore partially shows the effectiveness of our ranking system.

Finally, we also quantify the effect of the number of topics and its added amount to each webpage and ad by testing with different topic models and adjusting the *scale* value. However, it shows that the effect of the number of hidden topics does not cause a significant change in overall performance. As indicated in table 3, the performance of 120 and 200-topic models yield a better result than 60-topic model. However, there is no considerable change between 120-topic and 200-topic models, also in the quantities of added topics to each page and ad message. It can therefore conclude that the number of topics should be large enough to discriminate the difference of terms to better analyze topics for web pages and ads. And since the number of topics is large enough, the performance of the overall system is quite stable.

5.5. Chapter Summary

In this chapter, we have presented our variety experiments in contextual matching using 100 web pages and 2,706 ad messages. We carefully evaluated the performance of each strategy by examining 100 web pages and their corresponding proposed ad ranked lists. We used 11-point average precision and calculated the corrected ad messages found in top three positions of each ranked list. The result has pointed out that keywords are an important factor in contextual matching.

We then examine our approach using hidden topics by comparing them with the better previous method. It has shown a significant improvement in matching and ranking. Our system has successfully overcome the miss-match because of unimportant words and words that have different meanings (homonym) by taking into account their topic analysis.

Chapter 6: Conclusions

Online advertising in general and contextual advertising in particular are new and potential fields for researchers to study and promised to be a new trend in the economy of Vietnam. It is expected to grow very fast in the next few years. Matching and ranking ad messages in order to display them efficiently are an important part of this system and have attracted a lot of controversies in information retrieval community lately. However, literature publications in this field are still very sparse. The main objective of our thesis is improving the performance of matching system in order to propose relevant ads for a corresponding web page. In this chapter, we will summarize and conclude our main contributions as well as the future works in this area.

6.1. Achievements and Remaining Issues

In this thesis, we have given an overview of online advertising and investigated the problem of matching and ranking in contextual advertising. We have introduced some strategies that have been applied for solving this problem recently.

We then propose a new framework to investigate this problem using hidden topics model. This new approach has shown its high efficiency through a variety of experiments against the basic method using syntactic information only. It can overcome the problem of miss-match by discovering the latent semantic relations of web pages and ad messages and expanding their vocabularies. In practical, the results record an error reduction of 22.9 percent in the method HT200_20 over the matching strategy without hidden topics AD_KW. Further more, it indicates that this high quality contextual advertising framework is feasible and practical in reality. It has simply taken advantages of large scale external datasets that are available in the internet and not difficult to collect. Moreover, our framework is also flexible and general enough to be applied in a multilingual environment.

Our work described in this thesis has only focused on the problem of matching web pages and ad messages to suggest most relevant contextual ads. To optimize the revenue from contextual advertising, this framework also has to take into account features from an economic model, such as the keyword bid information of advertisers.

6.2. Future Work

Other areas of future work include applying these techniques to multimedia advertising, finding a ranking function to better describe the contribution of other features for finding most relevant ads (such as structural information, keyword bid information, etc.). Moreover, we want to examine more in details the effect of each topic model to optimize its efficiency, such as the number of topics in models and the best *scale* to balance the number of added hidden topics to web pages and ads. The ultimate goal of our work is building an advertising network in Vietnam based on this model with the purpose of improving the performance of online advertising, which is a promising and potential area in Vietnam though has not been studied and applied widely yet.

Vietnamese References

[1] http://24h.com.vn: 24H online advertising JSC.

[2] Bộ Thương Mại. Báo cáo thương mại điện tử Việt Nam 2006, 01/2007, http://www.mot.gov.vn.

[3] DanTri e-newspaper, DanTri.com.

[4] http://english.vietnamnet.vn/biz/2006/09/609892/: Online advertising potential in Vietnam (08/09/2006).

[5] http://www.vietad.com: Vietnam Advertising.

[6] http://vaa.org.vn: Vietnam Advertising Association VAA.

[7] http://www.quangbaweb.com/chienluoc.htm: Vinalink Media.

[8] http://VnExpress.net/: VnExpress: An Online Vietnamese news.

[9] Vietnamese Zing directory, http://directory.zing.vn/directory, 2008.

[10] http://VietnamNet.vn: VietnamNet e-newspaper.

English References

[11] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, Berthier A. Ribeiro-Neto: Learning to advertise. SIGIR 2006: 549-556

[12] Andrei Z. Broder, Marcus Fontoura, Vanja Josifovski, Lance Riedel: A semantic approach to contextual advertising. SIGIR 2007: 559-566

[13] B.Ribeiro-Neto, M.Cristo, P.B.Golgher, and E.S. de Moura. Impedance Coupling in Content-targeted Advertising. In *SIGIR 2005:*496-503, New York, NY, 2005.

[14] http://www.ciaadvertising.org: CIA Advertising

[15] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. In *Journal of Machine Learning Research:* 993-1022, January 2003.

[16] G. Heinrich. Parameter Estimation for Text Analysis. *Technique report*. 2005.

[17] Girolami, Mark; Kaban, A. On an Equivalence between PLSI and LDA. SIGIR 2003: 433-434, 2003.

[18] Hofmann, T., Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*: 177-196, 2001.

[19] http://www.iab.net/insights_research/1357: IAB Internet Advertising Revenue Report conducted by PricewaterhouseCoopers (PWC), http://www.iab.net.

[20] http://www.archive.org: Internet Archive.

[21] http**://**e-sgh.pl/cia/TRENDS.rtf: Advertisng/Direct Marketing Trends**.**

[22] C.H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. In *Proocedings of the ACM Conference on Principles of Database Systems (PODS)I:* 159-168, Seattle, 1998 http://citeseer.ist.psu.edu/papadimitriou98latent.html

[23] M. Ciaramita, V. Murdock, and V. Plachouras. Semantic Associations for Contextual Advertising. In *Journal of Electronic Commerce Research: Special Issue on Online Advertising and Sponsored Search*, **9** (1): 1-15, 2008.

[24] http://www.msn.com/: Microsoft Social Network MSN.

[25] Nguyen Cam Tu, "JVnTextpro: A Java-based Vietnamese Text Processing Toolkit".

[26] Nguyen Cam Tu, "JGibbsLDA: A Java and Gibbs Sampling based Implementation of Latent Dirichlet Allocation (LDA)".

[27] Nguyen Cam Tu, Hidden Topic Discovery toward Classification and Clustering in Vietnamese Web Documents, *Master Thesis*, College of Technology, Vietnam National University, Hanoi, 2008.

[28] http://lucene.apache.org/nutch/: Nutch (an open-source search engine).

[29] http://www.onlineadvertising.net/ : Online Advertising: news and quality online advertising information.

[30] Phan Xuan Hieu, "JTextPro: A Java-based Text Processing Toolkit", http://jtextpro.sourceforge.net/, 2007.

[31] Phan Xuan Hieu, "GibbsLDA++: A C/C++ and Gibbs Sampling based Implementation of Latent Dirichlet Allocation (LDA)", http://gibbslda.sourceforge.net/, 2007.

[32] Phan Xuan Hieu, Susumu Horiguchi, Nguyen Le Minh. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *17th International World Wide Web Conference*, 2008.

[33] T. Hofmann. Probabilistic LSA. *Proc. UAI*, 1999.

[34] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1):1-47, 2002.

[35] T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds). Latent Semantic Analysis: A Road to Meaning. *Laurence Erlbaum*, 2005.

[36] G. Salton, A. Wong, C.S. Yang. A Vector Space Model for Automatic Indexing, *Communication of the ACM*, **18** (11), 1975.

[37] C. Wang, P.Zhang, R. Choi, and M. D. Eredita. Understanding consumers attitude toward advertising. In *Eight Americas conf. on Information System:*1143-1148, 2002.

[38] Wen-tau Yih, Joshua Goodman, Vitor R. Carvalho: Finding advertising keywords on web pages. WWW 2006: 213-222.

[39] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*: 69-90, 1999.