Introduction to

KarnıYarık .com

Vertical Search Engine for Online Goods

1 What is KarnıYarık?

KarnıYarık is a *vertical search engine* specialized in product search. Its main purpose is to guide both online and offline shoppers (potential customers) about products of various types. It not only supplies information about product prices but also brands and categories. KarnıYarık utilizes the simplicity and power of *Free Text Search* (Google way). As we believe and some researches show, this style of searching is critical for *Value Shopping*. Our ultimate goal is to achieve nearly-full automation with more powerful *Faceted Search* capabilities and collected product reviews.

The architecture of KarnıYarık is based on *Web Crawlers* and *Semi-Automatic Parsers* meaning it does not require "data feeding" for data collection. There are numerous concepts and techniques that are incorporated (or planned to be incorporated) in the system such as *Natural Language Processing*, *Artificial Intelligence*, *Machine Learning*, *Social Networks*, and *Semantic Web*. One of the most advantageous properties of KarnıYarık is that its architecture is almost independent of product search domain and that allows us to have a generic vertical search engine framework which can be applied to various other domains such as used-cars, real-estate, news, etc.

A prototype version of KarnıYarık is running on a dedicated server located in Canada. Right now, it is capable of indexing over half a million products (pages) among 50 Turkish online shopping sites. But thanks to its proprietary architecture, reaching multiples of these values are just a matter of time and money (need more servers).

2 General Architecture

The project output will be a vertical search engine specialized in online shopping domain. Vertical search is a specialized version of Internet Search, where the search is applied on a particular domain or information type. Target of Vertical Search Engine is a set of audiences and collect information which audience will be interested.

Architecture of a Vertical Search Engine

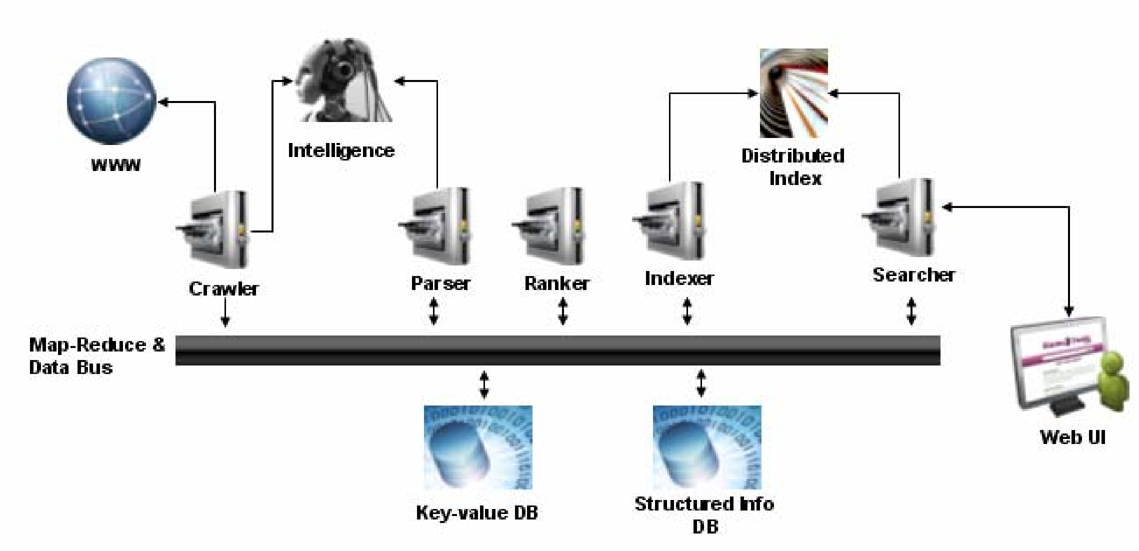


Figure 1 Architecture of a Vertical Search Engine

The architecture of a Vertical Search Engine is very much like a general purpose search engine. The main difference lies in the component named as "Intelligence" in **Figure 1**. This component chooses the Internet pages that are specialized to the selected domain, and extracts information that will attract the target audience. The other components are *crawler*, *parser*, *ranker*, *indexer* and *searcher*. The interactions between components are given in **Figure 1**. In order to develop these components technologies like *Artificial Intelligence*, *Information Retrieval*, *Information Extraction*, *Natural Language Processing*, and *Concurrent Programming* are required. Each component is described briefly in the following paragraphs.

2.1 *Crawler*

Crawlers visit the internet pages and try to identify the pages that are related to the focused domain. They download the pages, analyze their content and URL, and mark them as the focused pages. The main difficulties for crawlers are:

- The number of pages in Internet
- The change frequency of the pages
- Dynamically generated pages

In order to handle such difficulties there are some policies that a crawler must be careful about:

- *Selection Policy*: Deciding which pages to be downloaded and analyzed further.
- *Revisit Policy*: Deciding the time intervals to revisit and download a page.

- *Politeness Policy*: Since the crawlers are programs that can make many requests in seconds, they might cause a denial of service on the server that serves a site. In order to prevent this, a crawler must be polite when requesting pages from the server.
- *Parallelization Policy*: The huge number of pages in the internet, and the size of data to download require a parallelization in order to divide the jobs.

Despite the level of complexity they bring into the system, these policies must be taken into account when developing a crawler.

2.2 *Ranker*

This component calculates the importance of each page marked by taking the in/out links of a page and the page content into account.

2.3 *Parser*

Parsers take the marked pages as input. These pages are marked as related with the focused domain by the crawler. Parsers try to extract information from these unstructured HTML pages and output structured information. The extraction process can be manually programmed; or automatically accomplished with help of some *Artificial Intelligence* and *Statistical Methods*. *Machine Learning* is also used to recognize the information from past experiences; however, such systems require heavy training phases. While searching for online goods; product names, prices, product properties, brands, model numbers are some of the concepts that need to be extracted from the marked page.

2.4 *Indexer*

Indexers use *Information Retrieval* methods to locate the searched information in a short time period. *Tree structures*, *N-Gram*'s and *Inverted Matrices* are some type of core methods to index a given set of information. The *Inverted Matrix* is currently a popular way of indexing. In order to construct an *Inverted Matrix* the words in the given document has to extracted, stemmed, processed according to the domain (the output is called as "Term"). Finally a lookup table is constructed where each row represents a term, and each column represents the documents that the term or that row is included.

2.5 *Searcher*

Searcher component handles the querying of collected and indexed information. The Searcher processes the user query, extracts some Terms and rules, performs some calculations over Terms matched; and finally returns the results and scores of each result.

2.6 *Other components*

2.6.1 Sociality

This component enables users to construct a social web while using the system. The social web is used to return search results that the user might be likely interested using the profile and the connections of the user. Moreover it is used for recommending some other information in the system.

2.6.2 Recommendation System

Recommendation systems are used to make users reach the information they search but are not aware of. These systems are trained with some rules during development phase. Furthermore, it analyzes users' actions and come up with new additional rules according to these analyses. They finally recommend some information which is not searched but might be related to the current search.

2.6.3 Semantic Web

Semantic Web uses ontologies in order to annotate the information. When these annotated information is presented to the users via an API provided or via Web Pages, it becomes possible to recognize the meaning of some words or word groups in the information. Our project will be using *Microformat* to annotate the product information presented to users. Moreover, we plan to open an API to the public which will expose the knowledge in the system. Exposing annotated data will enable API users to easily understand the information, and will save them from the complex tasks that our project accomplishes.

3 Motivation

3.1 *Technological Vision of Turkey*

Vision 2023 is a Turkish project developed to make use of science and technology as much as possible on the way of becoming a European country. As a subset of this large project *Information and Communication Technologies* panel is prepared. Contents of this panel clearly show how KarnıYarık fits to Turkish vision of *Information and Communication Technologies*.

The panel states that *Artificial Intelligence*, *Natural Language Processing* and *Search Engines* are technologies that will evolve and be widely used in near future. To automatically investigate and extract information from *World Wide Web* KarnıYarık plans to make use of *Artificial Intelligence*, *Natural Language Processing* and *Semantic Web* technologies where possible.

Again according to the panel following technologies are indicated as of first priority;

- Artificial Intelligence
- Web Programming

- Data Mining
- Natural Language Processing
- Interoperable Software Architectures
- Fuzzy Logic
- Distributed Systems and Parallel Programming

As it is described in project definition and architecture sections KarniYarik will serve to its customers using all those technologies.

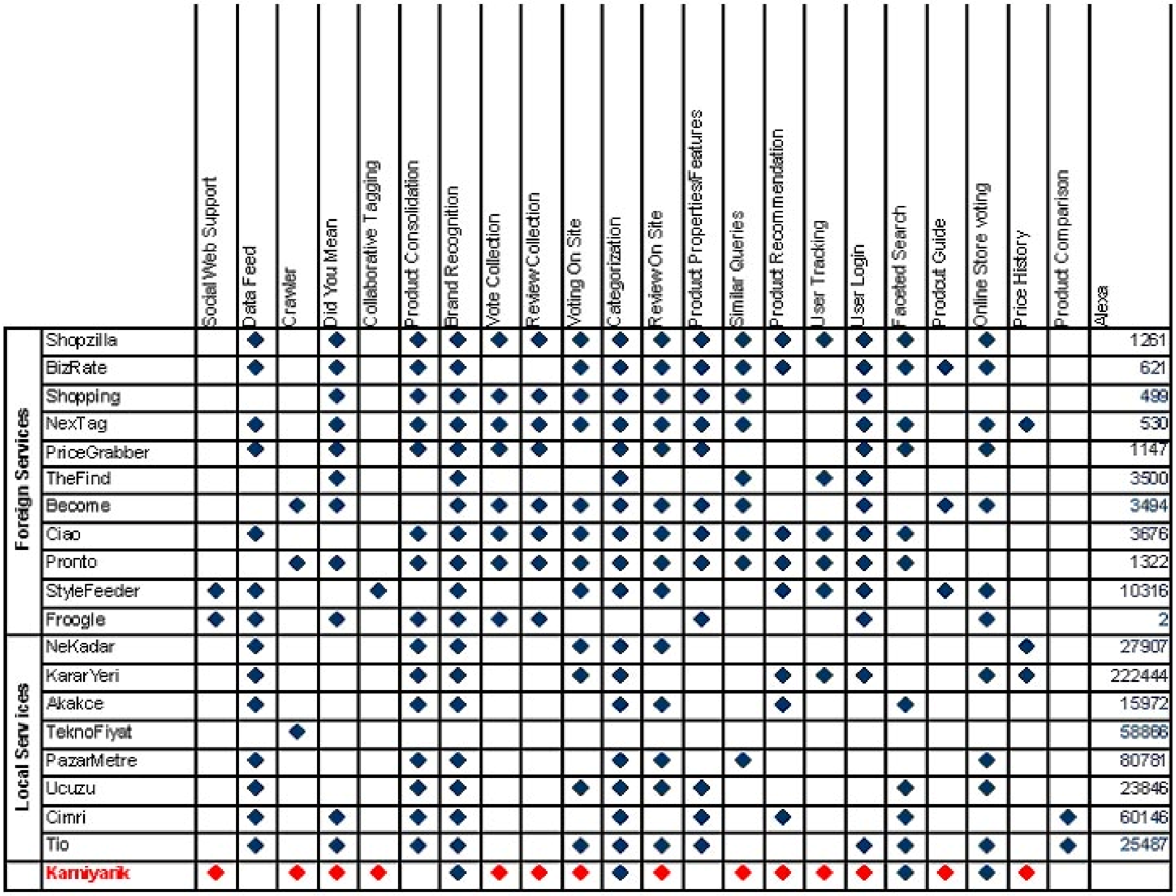
3.2 Reasons behind KarnıYarık

Today, searching the Web and helping people to find information has become one of the largest business areas. Increasing number of internet users makes this area more and more suitable for entrepreneur projects. On the other hand, the need to find correct information within minimum time, has evolved so much that classical search engines are can't solve this problem very efficiently. To serve internet users in a better way, Semantic and Vertical search engines are being developed as an alternative to general purpose search engines. Today's well known search companies are also making researches on these subjects, especially on vertical searching. Guided by these facts and motivated by our passion to achieve worldwide success on this evolving business area, KarnıYarık will become a Turkish vertical search engine.

Turkish online shopping is KarnıYarık's first target area. Online shopping is chosen because current Turkish search engines for this area do not provide some services that are provided by their foreign conjugates. And manual data collection processes that are currently used can be replaces by automatic data collection technologies.

Even though Turkish online shopping is our first target area, extending our business to English online shopping on other countries and then to other vertical search subjects with the experience developed is our future plan.

3.3 Functional Matrix



		Social Web Support	Data Feed	Crawler	Did You Mean	Collaborative Tagging	Product Consolidation	Brand Recognition	Vote Collection	Review Collection	Voting On Site	Categorization	Review On Site	Product Properties/Features	Similar Queries	Product Recommendation	User Tracking	User Login	Faceted Search	Product Guide	Online Store voting	Price History	Product Comparison	Alexa
Foreign Services	Shopzilla		◆		◆		◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆		◆			1261
	BizRate		◆		◆		◆	◆			◆	◆	◆	◆	◆	◆		◆	◆	◆	◆			621
	Shopping				◆		◆	◆	◆	◆	◆	◆	◆	◆	◆			◆						499
	NexTag		◆		◆		◆	◆	◆	◆	◆	◆	◆	◆	◆			◆	◆		◆	◆		530
	PriceGrabber		◆		◆		◆	◆	◆	◆		◆	◆	◆				◆	◆		◆			1147
	TheFind				◆			◆				◆			◆		◆	◆						3500
	Become			◆	◆			◆	◆	◆	◆	◆	◆	◆	◆			◆		◆	◆			3494
	Ciao		◆				◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆					3676
	Pronto			◆	◆		◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆	◆					1322
	StyleFeeder	◆	◆			◆		◆			◆	◆	◆			◆	◆	◆		◆	◆			10316
	Froogle	◆	◆		◆		◆	◆	◆	◆				◆				◆			◆			2
Local Serv ices	NeKadar		◆				◆	◆			◆	◆	◆									◆		27907
	KararYeri		◆				◆	◆			◆	◆				◆	◆	◆			◆	◆		222444
	Akakce		◆				◆	◆				◆	◆			◆			◆					15972
	TeknoFiyat			◆																				58866
	PazarMetre		◆				◆	◆				◆	◆		◆						◆			80781
	Ucuzu		◆				◆	◆			◆	◆	◆	◆					◆		◆			23846
	Cimri		◆		◆		◆	◆				◆		◆		◆			◆				◆	60146
	Tio		◆		◆		◆	◆			◆	◆	◆	◆				◆	◆		◆		◆	25487
	Karniyarik	♦		♦	♦	♦		◆	♦	♦	♦	◆	♦		♦	♦	♦	♦	◆	♦	◆	♦		

Figure 2 Service/Functionality Matrix

Figure 2 displays a list of features that are serviced for online shopping and a list of Turkish and foreign websites providing vertical search services for online shopping. Features supported by each website are marked with a blue color. Features that will be supported by KarnıYarık are marked with red color.

KarnıYarık will especially be superior to its local competitors by automating the data collection. Those service providers which manually collect information from online shopping sites suffer from large number of online shopping sites and high frequency change rate of online product data. To eliminate these problems KarnıYarık will search online shopping centers with a specialized crawler to find important shopping information. There is only one shopping service using this technique, and there are a few foreign services that use automatic crawlers to collect data.

One of the first improvements that will be implemented is social search service. This functionality is not provided by any Turkish vertical search engines. We are planning to provide search results based on peoples interests and friends. Same query would return different results for different persons with respect to their social information.

To provide better search results KarnıYarık will make use of *Knowledge of the Crowd* concept. This functionality will be used to categorize and to grade search results. Users will be able to tag product information provided to them as search results. Information collected from users will be used together with our own categorization. This functionality is not present in Turkey, and it is provided by only a few number of foreign vertical search engines.

Another functionality which is not present in Turkish market is collecting product comments from forums and other discussion sites. KarnıYarık will collect discussions of users on products and display this information in its search results. Later on guidelines for product models will be composed, by which users will learn critical properties of products.

Recommendation engine for search queries, which is a must for search engines, will be implemented. Even though most of foreign search engines provide this functionality, it is not common on Turkish market. By this functionality on misspelled search queries users will be guided.

And finally Semantic search functionality which is not given in **Figure 1** is a research area for us. Semantic searching is a new upcoming concept and integrating it to vertical search engines is a brand new concept.

After explaining all those features that KarnıYarık will implement, we should sum up with our general goal. Implementing a vertical search engine for Turkish Online Shopping will provide us a huge experience. The basic idea is to develop a Crawl and Search framework which will be used for future vertical search projects.

3.4 Market Analysis

As given in **Figure 1** KarnıYarık's competitors on Turkish Online Shopping market are *NeKadar.com*, *Akakce.com*, *Ucuzu.com*, *Cimri.com*, *Pazarmetre.com*, *KararYeri.com* and *Teknofiyat.com*. Most widely used service within those is *Akakce.com*, which is ranked 15972th most visited website in the world by *Alexa*. On the other hand the most widely used Online Shopping website in Turkey is *Hepsiburada.com*, which is ranked 1749th most visited website in the world. Turkish online shopping market is KarnıYarık's first target and we plan to become the gateway of users that will shop online or investigate any online good. In other terms, our first indicator of success is to become more popular than *Hepsiburada.com*. Such a success is only possible with high quality services and good advertising policy. Target users for KarnıYarık are people who;

- wants to get information about a product
- wants to know where to buy a product
- wants to know prices of a product provided by several shopping sites
- wants to see discussions and comments of a product

Turkey currently has 26 M Internet users which places the country in the 11th position in the world Internet usage(Internet World Statisctics). According to the presentation titled *"Turkish Internet Sector Overview"* by *Sina Afra (eBay, Germany)* the importance of gateways for that amount of users is increasing. These gateways are used for email, search, shopping, news, weather, maps an etc. Moreover, it presents that they expects introduction of advenced advertising technologies and international partnerships in the advertisign segment.

Within last few years online shopping has been increasing in the world market, and it is also increasing in Turkey. According to BKM Internet Card Reports trading volume in Turkey over virtual POS machines has been constantly increasing since 2005 and reached to 9M $ by 2008. Again according to information in http://www.sanalmimarlar.com/images/etic2008-2Ceyrek.pdf this amount is only %5 percent of shopping made by credit cards. Which means online shopping will is expected to get bigger in Turkey. KarnıYarık will guide aprroximately 30M users in 9M $ worth market.

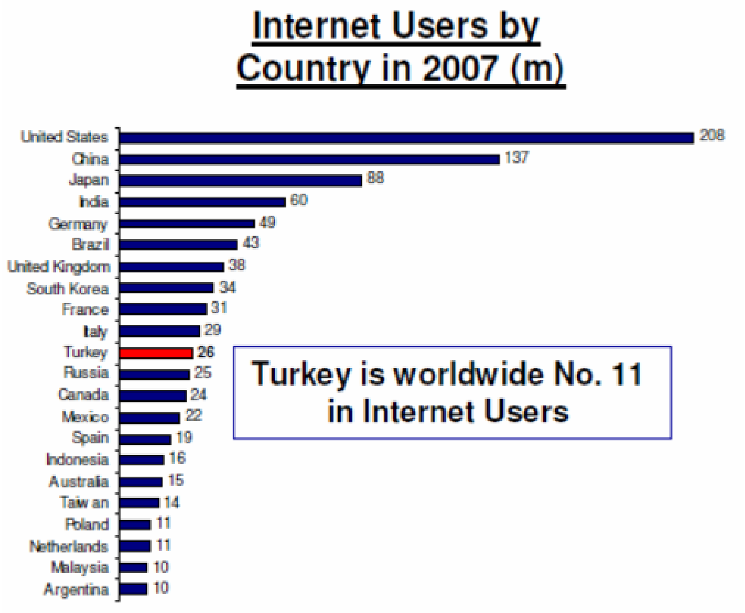


Figure 3 World Internet user statistics

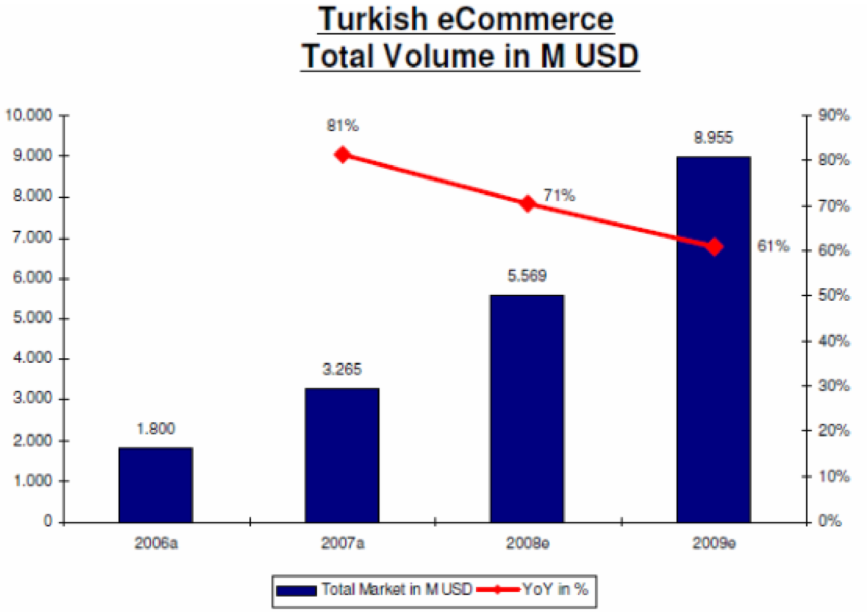


Figure 4, E-Commerce transactions made over virtual POS machines

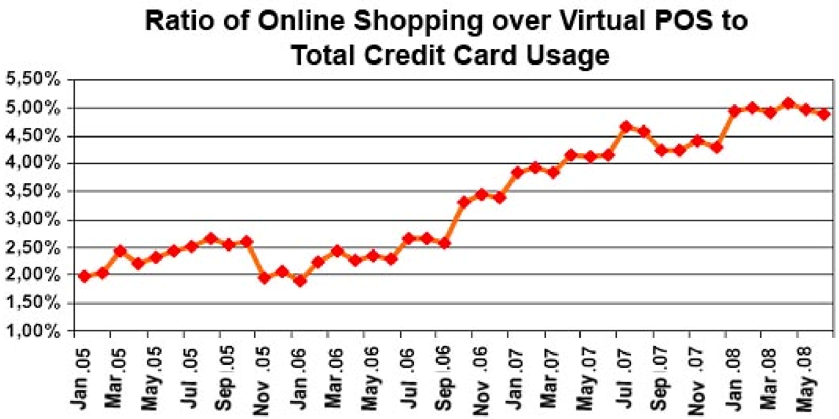


Figure 5, Percentage of virtual POS transactions to Credit Card transactions for shopping

Finally, another market that KarnıYarık will serve is the online advertising market that has 46M $ value in Turkey in 2007. KarnıYarık expects to share a considerable amount of online advertising market which is based on online goods.

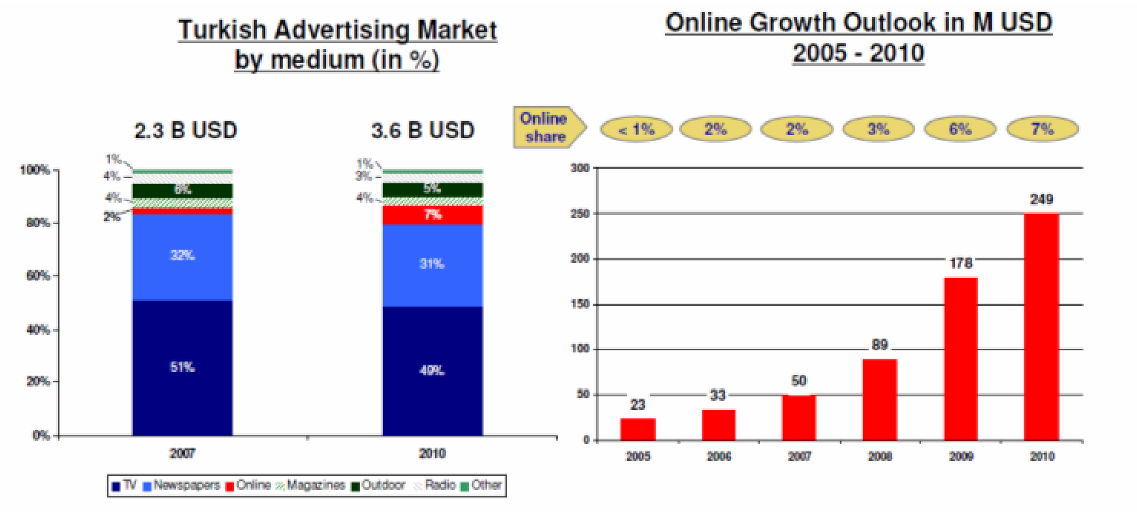
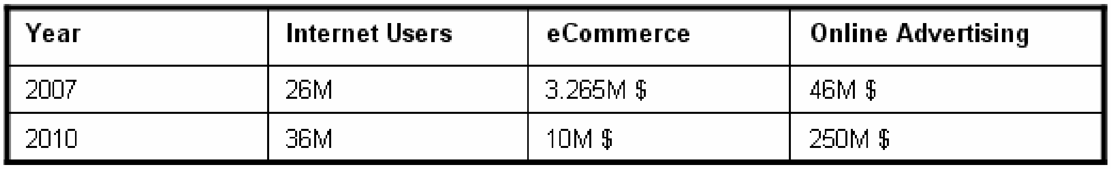


Figure 6 Advertising Market in Turkey



Year	Internet Users	eCommerce	Online Advertising
2007	26M	3.265M $	46M $
2010	36M	10M $	250M $

Figure 7 Summary of current value of target markets in 2007 and expectations in 2010.

4 Possible Business Model

We do not have any professionalism in internet business so honestly, the ways of monetizing KarnıYarık is a little bit fuzzy in our minds. One of the well-known techniques is online advertisements and we are planning to have sponsored links for products and shopping markets. KarnıYarık will collect various statistics about online shopping market and we think this information is also very valuable. Besides we will construct a content-based advertising system inside KarnıYarık.

Being one of the first talented groups working on the technologies involved in KarnıYarık like web crawling, information extraction or search, in Turkey is like a free ticket to consult or directly take action in different military and governmental projects.

Last but not least, as we mentioned earlier, KarnıYarık is developed on an architecture which is nearly independent of online shopping domain which brings us the opportunity of applying it to different domains such as real-estate, news, used-cars, etc.

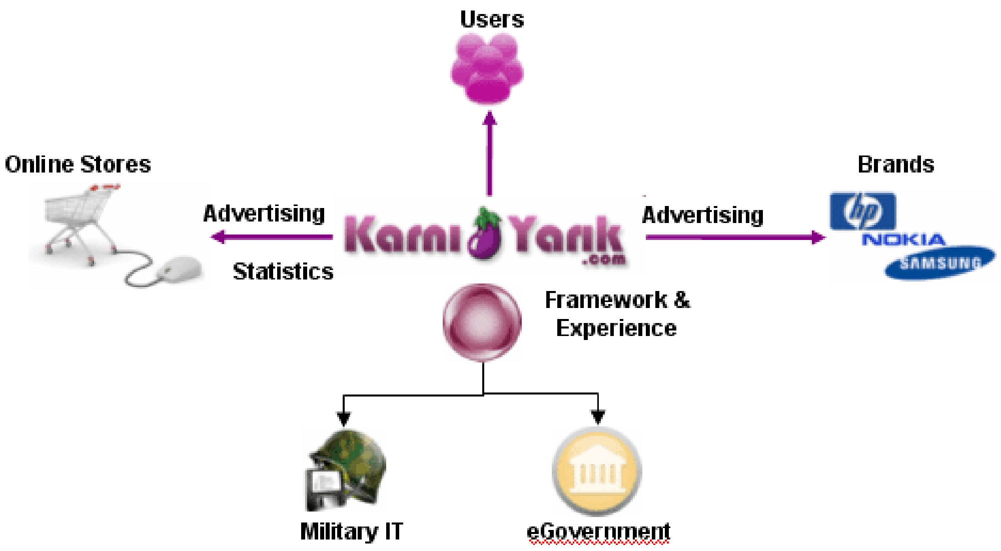


Figure 8 The marketing ecosystem

5 Conclusion

We tried to explain what KarnıYarık is and present how we can make money out of it. Although we are not crystal clear on monetizing methods, we are sure that if we can make KarnıYarık popular in Turkey, we can also make considerable profit. We also believe that we are a very talented and strong team that will be distinguished on the technologies we will be working on. Eventhough Turkey is our first target country, once KarnıYarık is up and running with a stable architecture, it will be very easy to extend our services to foreign countries. In order to realize these things, we need guidance and funding by experienced investors.

6 Glossary

Free Text Search: In text retrieval, full text search refers to a technique for searching a computer-stored document or database. In a full text search, the search engine examines all of the words in every stored document as it tries to match search words supplied by the user.

Value Shopping: Shopping effectively like by finding the cheapest and best quality product or service.

Faceted Search: A technique for accessing a collection of information represented using a faceted classification, allowing users to explore by filtering available information.

Natural Language Processing: A field of computer science concerned with the interactions between computers and human (natural) languages.

Social Networks: A social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friendship, kinship, dislike, conflict or trade.

Semantic Web: It is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content

Concurrent Programming: A form of computing in which programs are designed as collections of interacting computational processes that may be executed in parallel.

N-Gram: A type of probabilistic model for predicting the next item in a sequence.