ENPRO 356: Arabic Stemmer

النوئ انستيتيوت آف تكنولوجي

(Illinois Institute of Technology)

Spring 2003 Team

Faculty Advisor		Dr. David Grossman
Project Leader		Michael Lee
Student Leaders		Steve Beitzel Eric Jensen
Members	Syed Aqeel Chirag Bhatt Raghu Kutty Fan Ping Nayana Samaranayake	Aaron Samuels Palak Shah Sterling Stein Mudit Tandon George Taylor

Business Organization and Key Objectives

Sales	Marketing	Support	Product Dev.	Finance
▪Personal demo for Convera and their customers ▪Word of mouth advertising from Convera ▪Possible public ads	▪Establish deal with Convera ▪Advertise the strength of our stemmer ▪Form relations with other search engine companies	▪Create helpdesk or other support line for customers ▪Keep close contact with Convera in case of any problems found	▪Develop plug-in for other search engines ▪Improve and increase lexicon ▪Develop other language stemmers	▪Limit expenses upon start-up ▪Try to maintain free-cash-flow ▪Achieve break-even point as early as possible

Search Engines

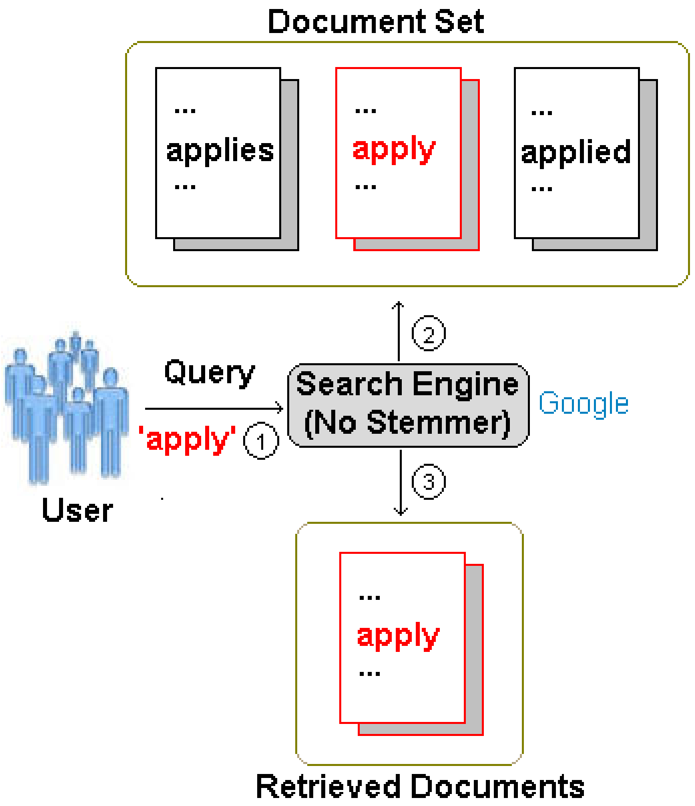
- Search engines find specific information in large volumes of data
  - User submits **query**, or what to search for
  - Search engine searches its **document set** for the query
  - Returns formatted results to the user
- Search engines are the most visited websites
  - Google
  - Yahoo!

Stemmers

- Stemmers group words with a common root together
    - [hypnotize, hypnotist, hypnotics] reduced to the same stem
- Improves accuracy because groups are of words with similar meanings
- Our stemmer makes search engines handle Arabic more effectively

Role of A Stemmer

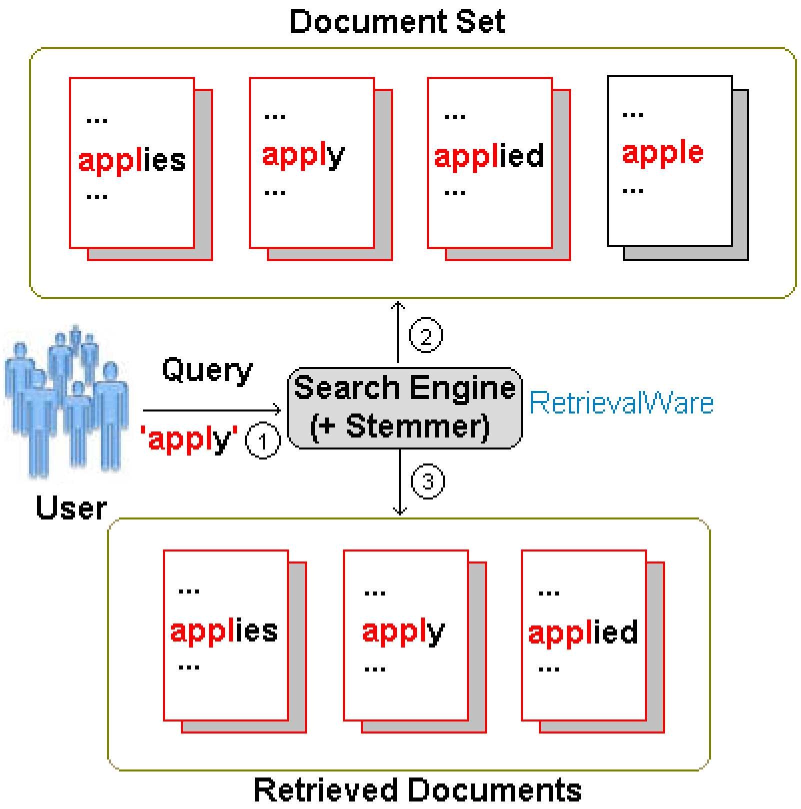
No Stemmer



Outcome:

Only 1 out 3 documents retrieved.

With A Stemmer



Outcome:

All 3 relevant documents retrieved!

Product

- The IIT Arabic Light Stemmer
  - Based on Mohammed Aljlayl’s award winning light stemming algorithm
  - Heavily optimized for efficiency
  - More effective than every comparable stemmer

Evaluation

- Participated in annual Text Retrieval Conference (TREC) in 2001 and 2002
  - Adjudged second best in the world in 2001
- Performance tested using TREC-2001 and TREC-2002 document, query, and result sets

Market Size & Target Buyers

- Primary Target: American Enterprises / Government Organizations that deal with the Middle-East
  - Resources aplenty in this field:
    - Oil / Gas companies
      - Halliburton Inc.
      - Shell
    - Government Agencies dealing with Middle-East issues
      - US Air Force
      - German Foreign Affairs Ministry
    - Multinational Corporations with operations in the Middle-East
      - McDonalds
      - Kraft

The User Value Proposition

- An effective stemmer makes applicable documents rank higher
  - Returns more accurate results
  - User finds desired information sooner
- An efficient stemmer can do the same work with fewer resources at a lower cost

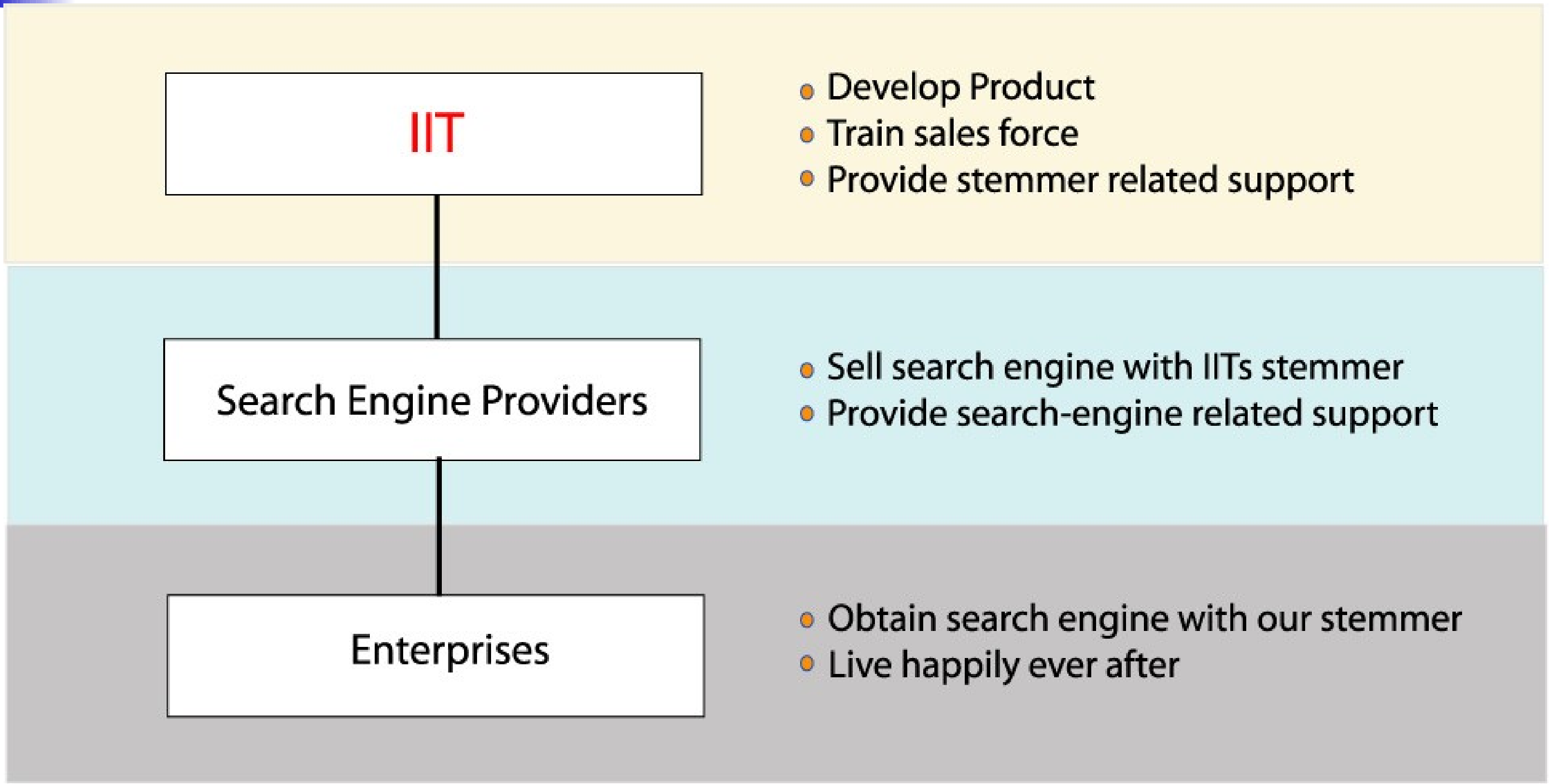
The Enterprise Value Proposition

- Cost savings
  - Saves valuable employee time
    - 5 min/wk x 500 employees x 52 wks = 4333 hrs
    - 4333 hrs x $20/hr = **$86,660 yearly**
- Convenience
  - Easy integration
  - Immediate Product Support

Search Utility Value Proposition

- Benefits to Search Utility Provider
  - Plug-in already built and integrated
  - More effective than their current solution
  - Improved and more marketable product

Supply Chain / Distribution



IIT

- Develop Product
- Train sales force
- Provide stemmer related support

Search Engine Providers

- Sell search engine with IITs stemmer
- Provide search-engine related support

Enterprises

- Obtain search engine with our stemmer
- Live happily ever after

Convera – Our First Partner

- More than 800 customers in over 30 countries
- 200 government customers; 80 involved in intelligence gathering
  - Customers such as the FBI, US Dept. of Defence, US Air Force
- Expressed interest in our product
- Makes search product: RetrievalWare

The RetrievalWare Search Engine

- Has the third largest market share in the enterprise search market (10.6%)
- Searches across more than 200 forms of text, video, image and audio information
- Searches performed over more than 45 languages
- *Allows for 3rd-party language plug-ins*

Our First 20 Orders

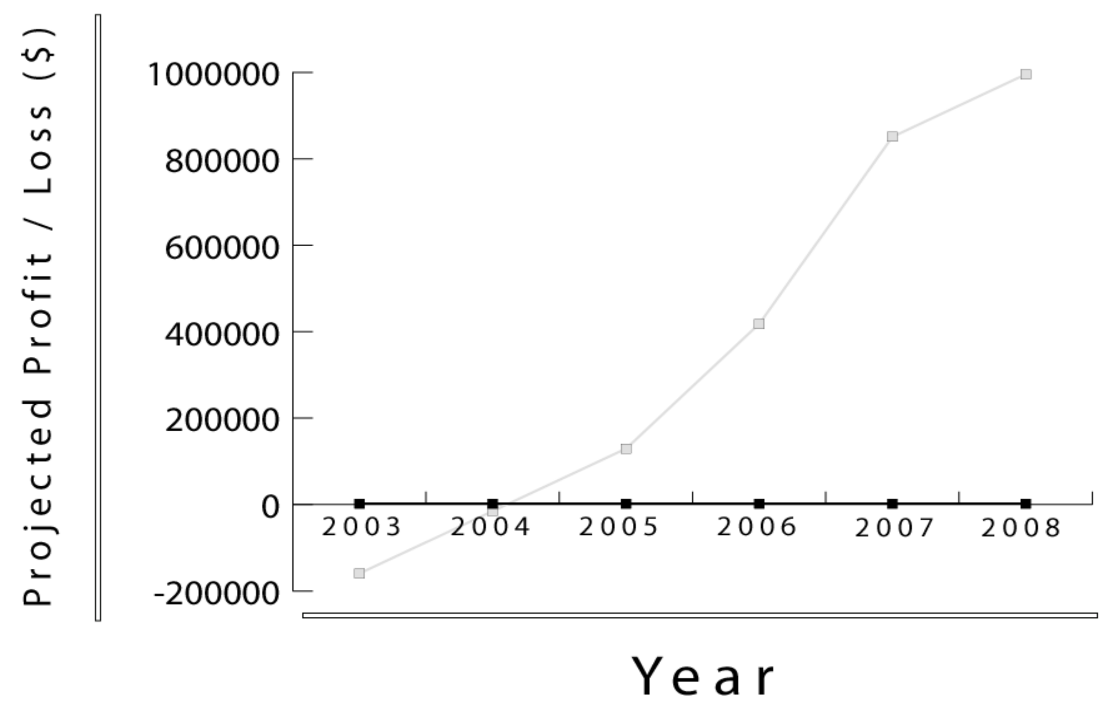
- Our first partner – Convera
  - Arrange demonstrations for the decision-makers of Convera
  - “Personal Marketing”
- Convera’s Customers
  - Already using RetrievalWare
  - Arrange demonstrations
  - Word-of-mouth advertising from Convera
- Next Step Out
  - Public advertising in the future to a wider audience

Pricing Model

Personnel Expenses:	$380,000
CEO, Programmer/Support, Sales, Others	
Operational Expenses:	$45,500
Rent, Utilities, Phone/Internet, Insurance	
Capital Expenditures:	$22,650
Computers/Network, Office supplies/maintanence	
Number of years we expect to break even in (Y):	3
Total Expenditure in Y years:	$1,299,150
Number of stemmers sold a year:	15
AVERAGE WE NEED PER STEMMER SOLD :	**$28,870**
Commission @ 10% per sale:	$3,208

AVERAGE PRICE PER STEMMER: $32,078

Cashflow Analysis



Top Three Direct Competitors

Competitor	Their Strengths	Their Weaknesses	Why We will Win!
Search Utility Providers’ existing utilities	Already have them – why pay someone else?	Not as effective as our stemmer	Proven unmatched effectiveness
Other Commercial Stemmers	Have a head-start	No international recognition	We have demonstrated that we lead at international conferences
Other Academic Stemmers	Comparable effectiveness to our stemmer	Completely academic version, not as easy to integrate to search tools	Ours is a commercial version. Ready to use by companies

Risks and Mitigation

- Company dependant upon one product
  - Expand into custom development
- Competition from other Arabic stemmers
  - We beat all existing Arabic stemmers

Path Forward

- Establish partnership with other key search engine developers
  - Verity
  - Autonomy
- Develop other Arabic specific IR tools
- Validate performance and gain recognition through international IR conferences

Summary

- Constant innovation has enabled us to outclass other Arabic stemmers
- Only academic stemmer that works with existing search utilities
- We will market our stemmer through strategic alliances with search utility providers
- We pitch our stemmer by showing its cost effectiveness and value to end users and companies

Thank You!

Questions?

Supporting Chart I:

Pricing Analysis

Pricing Analysis

Personnel Expenses

CEO	1	150,000	150000
Programmers / Support Staff	2	40,000	80000
Sales Staff	4	25,000	100000
Seceratary	1	20,000	20000
Cleaning Staff / Misc. Employees	3	10,000	30000
Total			380000

Operational Expenses

Rent Expense	12	2,200	26400
Utilities	1	1100	1100
Phone / Internet	12	200	2400
Misc. Expenses	12	300	3600
Small Business Insurance	12	1000	12000
Total			45500

Capital Expenditure

Computers	1500	6	9000
Laser Printer			500
Annual Maintanence Contract			1000
Network Router			150
Office Supplies			2000
Office Furniture			10000
Total			22650

Pricing Analysis

Number of years we expect to break even in (Y)		3
Total Expenditure in Y years		1299150
Number of stemmers sold a year		15
AVERAGE WE NEED PER STEMMER SOLD		**28870**
Commission @ 10% per sale		3207.778
AVERAGE PRICE PER STEMMER		**32078**

How our partners are priced

Vendor	Product	Pricing Arrangement
1. Convera	RetrievalWare	$75K(average)
2. Verity	K2 Enterprise	$100K+
3. Autonomy	IDOL Server	$360K(average)
4. Microsoft	SharePoint	$3K(per Server)
5. FastSearch	FAST Search	$100K+

Conclusion: The search engines offered by these companies vary in scope, functionality, and cost. The value of our stemmer will be different for each one. Therefore, we should adjust the charge for our stemmer accordingly.

Supporting Chart II:

Assessment of Search Engine Partners

Partner Analysis

Area:	Verity	Convera	Autonomy
Overall Attractiveness as a Partner	***	*	**
Customer and market - Strengths and challenges	Largest market shareholder of this industry. Leader of the business portal software infrastructure market	Established customer base. Sells to a lot of government agencies	Established over 130 partners worldwide. Recently became partner with China's No.1 IT manufacturer.
Product Technology – Strengths and challenges	Reliable and efficient knowledge retrieval technology	Very strong multi-lingual search capabilities	Leader of unstructured information retrieval. High product cost
Financial strengths/ Challenges	Strong financial position	Very low cash Dismal EBIDTA	Strong, positive cash flow with growing assets.