HEALTHKEE: A SYMPTOM CHECKER AND PERSONALIZED HEALTH INFORMATION SERVICE

Nguyen Van Giap, Le Dinh Minh, Vu Viet Anh
Advisors: *Dr. Phan Xuan Hieu*
VNU University of Engineering and Technology
Faculty of Information Technology

Introduction

With improvements in technology and access to the internet, people are increasingly using the Internet to research their health concerns. According to Pew Research Center's Internet & American Life Project, more than **35%** of adults in the United States regularly use the Internet to self diagnose their ailments, using it both for non-urgent symptoms and for urgent symptoms. There are many systems in English can support self-diagnosis as well as provide medical information such as WebMD, Mayo Clinic, NHS Choice, etc. However, there are not many same systems in Vietnam that can provide reliable and easy to understand clinical information. ***Healthkee*** aims to become a smart health assistant for Vietnamese that provide personalized health information, which is accurate, fast and easy to understand anywhere, anytime.

Medical data

- Collaborated with a group of students from **Hanoi Medical University.**
- **Provided by:**
  - Hanoi Medical University Hospital
  - Vietnam – Germany Hospital
  - Bach Mai Hospital
  - National Hospital of Obstetrics and Gynecology
  - Vietnam National Hospital of Pediatrics

Table 1: Medical concepts in database

Concept	#
Conditions	170
Symptoms	724
Risk factors	384
Medications	575
Tests and procedures	81
Complications	180

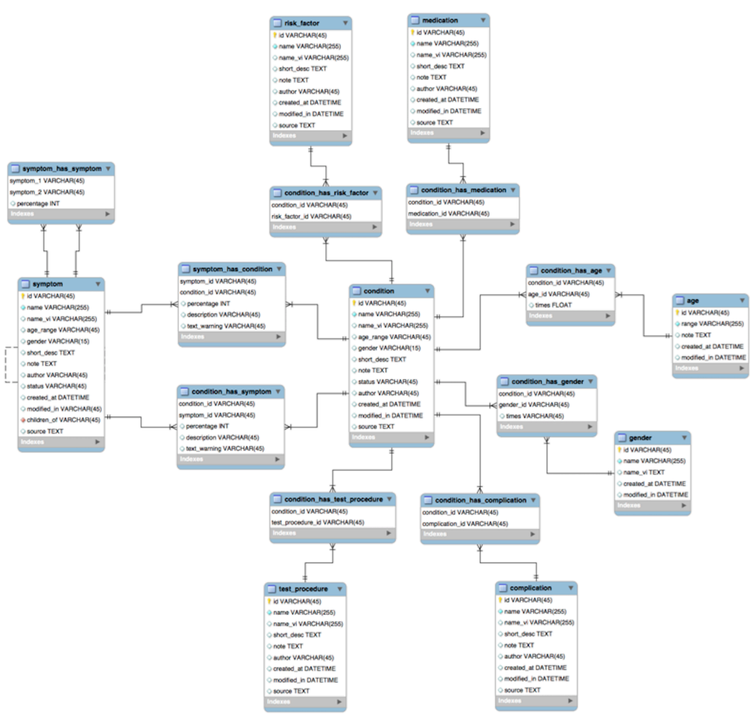

Figure 1: Database schema

Showcases

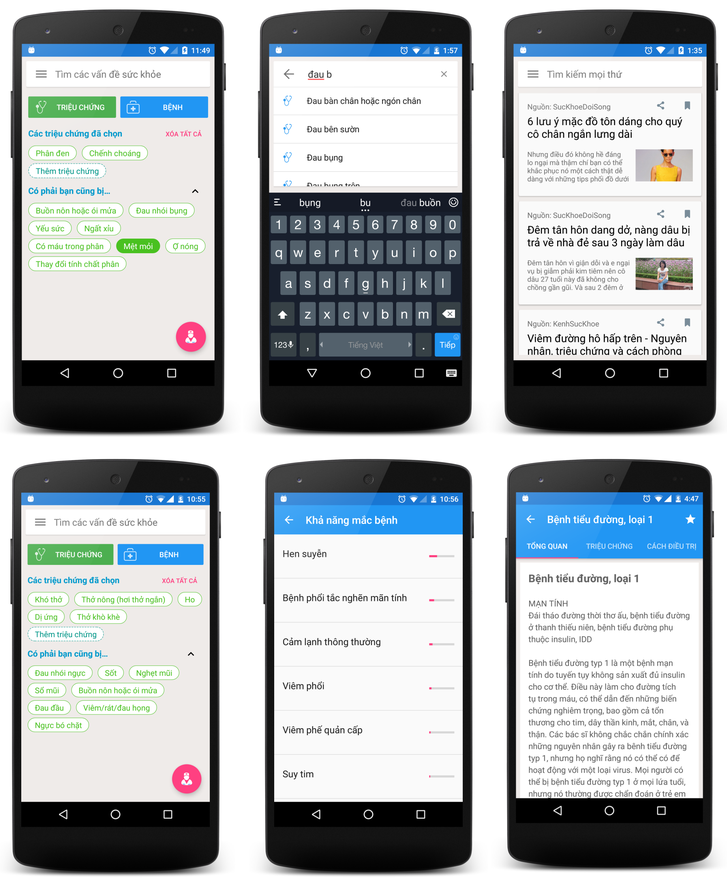

Figure 2: Structure of Search Module

Medical concepts searching

- Working **offline** on Mobile.
- Using **Inverted Index** technique.
- Searching with/without **accents** and **variants**
- Ranking used **element-wise comparison.**
- The relevance between user's query q and medical concepts c^i is a vector:

$$score(q, c^i) = \left(f(q_1, c^i), \dots, f(q_N, c^i), g(sl, c^i)\right)$$

in which

- $f(q^j, c^i) = sign(q^j, c^i) * s(q^j, c^i)$
- $g(sl, c^i) = sign(t^k, c^i) * s(t^k, c^i) + \min\{occur(sl, t^k)\} + length(t^k) - length(sl)$
- $s(q^j, c^i) = \min\left\{occur(q^j, c^i) | occur(q^j, c^i) \neq \underset{1 \le k \le j-1}{s}(q^k, c^i)\right\}$

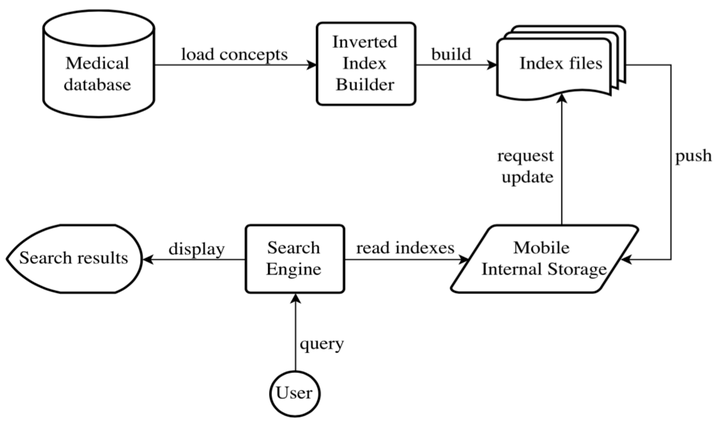

Figure 2: Structure of Search Module

Condition ranking

Find **all possible conditions** based on **chosen symptoms** and **demographics**, then rank them in an appropriate order: the most relevant conditions go first, the less relevant conditions go after or may be eliminated.

$$f(S_u, c) = \left(\sum_{s \in S_u \cap S_c} r(s, c)\right) * g(S_u, c) * h(s_u, c)$$

in which

- $|E_{uc}| = \{(s_i, s_j) | s_i, s_j \in S_{uc}, s_i \neq s_j\}$
- $g(s_u, c) = \left(1 + \frac{|S_u \cap S_c|}{|S_u|}\right) \log_2(1 + |S_u \cap S_c|)$
- $h(S_u, c) = 1 + \frac{\sum_{(s_i, s_j) \in E_{uc}} r(s_i, s_j) + r(s_j, s_i)}{2|E_{uc}|}$

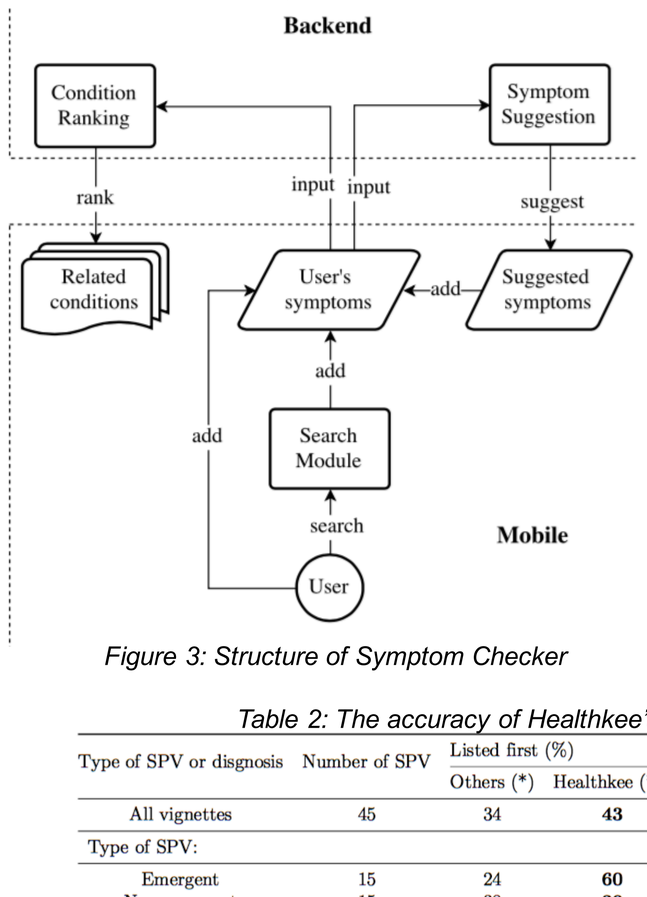

Figure 3: Structure of Symptom Checker

Symptom suggestion

- Suggesting **the most relevant symptoms** which users may concern but not selected and thereby help to lead users to appropriate diseases.
- Idea: find **all symptom candidates** then **rank** them by calculating the relevance between the set of symptoms chosen by users and those candidates:

$$score(s, S_u^t, S_{u_{NEG}}^t) = \log_2(1 + h) * \max_{1 \le i \le h}\left\{k^i \left(1 - e^{\alpha \omega_i^t}\right)\right\}$$

in which

- h number of conditions related to the candidate symptom s
- $k^i = |S_{c^i} \cap S_u^t|$
- $\omega_i^t = \frac{\sum_{s_j \in S_{c^i} \setminus S_{u_{NEG}}^t} r(s_j, c^i)}{\sum_{s_j \in S_{c^i}} r(s_j, c^i)} \in [0,1]$ with $r(s_j, c^i)$ demonstrates the relevance between s_j and c^i

Medical recommendation system

- Personalized distributing over **150K+ medical articles** from 20 well-known newspapers based on user interests.
- Using **Latent Dirichlet Allocation** to analyze hidden topics from articles.
- **Topic-sensitive recommendation** with latest user reading articles analysis.

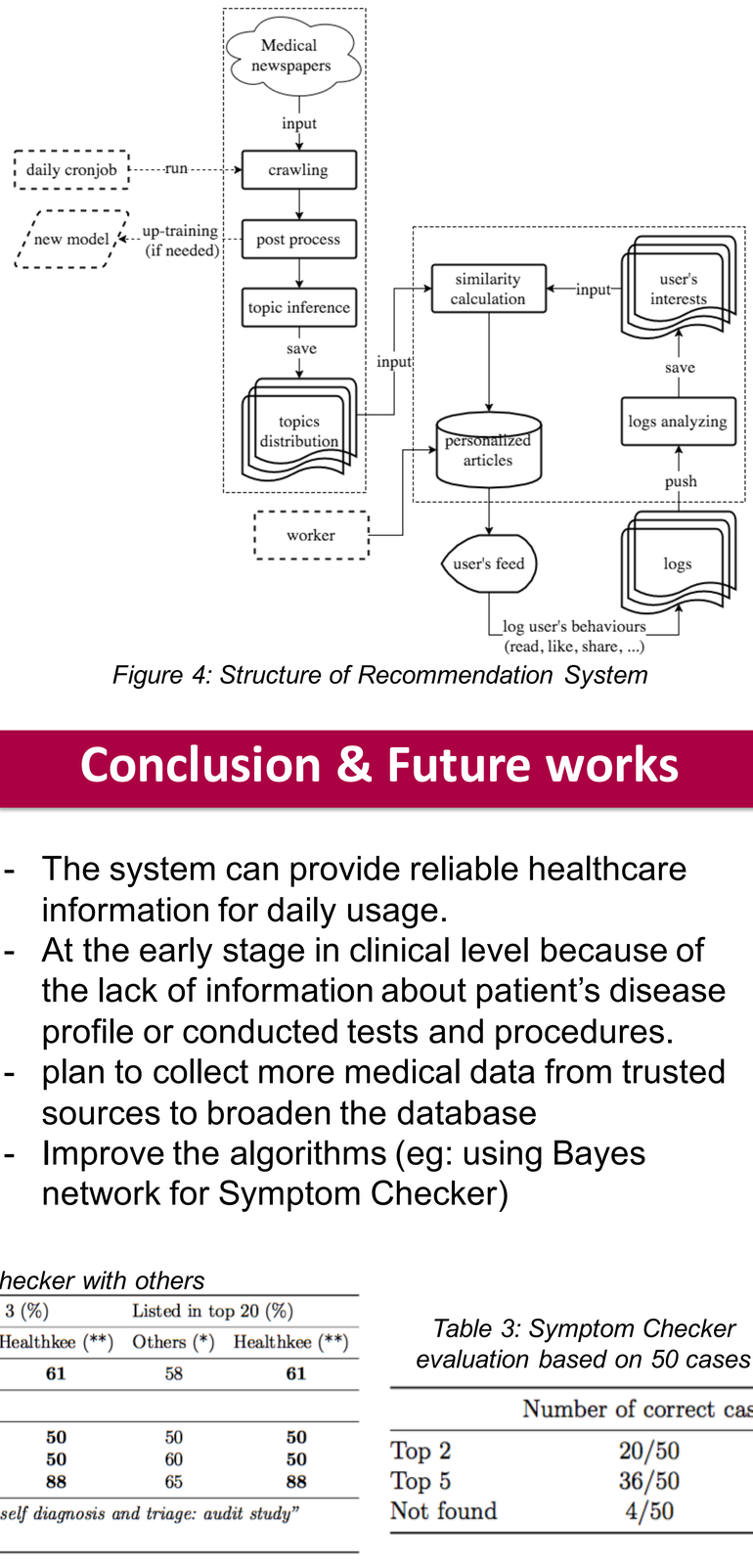

Figure 4: Structure of Recommendation System

Conclusion & Future works

- The system can provide reliable healthcare information for daily usage.
- At the early stage in clinical level because of the lack of information about patient's disease profile or conducted tests and procedures.
- plan to collect more medical data from trusted sources to broaden the database
- Improve the algorithms (eg: using Bayes network for Symptom Checker)

Table 2: The accuracy of Healthkee's symptom checker with others

Type of SPV or disgnosis	Number of SPV	Listed first (%)		Listed in top 3 (%)		Listed in top 20 (%)	
		Others (*)	Healthkee (**)	Others (*)	Healthkee (**)	Others (*)	Healthkee (**)
All vignettes	45	34	**43**	51	**61**	58	**61**
Type of SPV:							
Emergent	15	24	**60**	40	**50**	50	**50**
Non-emergent	15	38	**30**	57	**50**	60	**50**
Self care	15	40	**38**	57	**88**	65	**88**

(*): 23 symptom checkers mentioned in the study *"Evaluation of symptom checkers for self diagnosis and triage: audit study"*
(**): Number of correct SPV evaluations divided by applicable SPV evaluations.

Table 3: Symptom Checker evaluation based on 50 cases

	Number of correct cases
Top 2	20/50
Top 5	36/50
Not found	4/50