Toward an automated, self-sustaining knowledge base to facilitate linking point-of-sale grocery items to nutritional content

Valliammai Chidambaram, M.Sc; Philip J. Brewster, PhD; John F. Hurdle, MD PhD
Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Introduction
A public (open-source) electronic database linking nutritional information to grocery food item UPCs is currently not available for research use.
A data-sharing agreement with a national retail grocery store chain has allowed us to take steps towards the development of such a database.
Our previous research established the feasibility of linking supermarket sales data to relevant nutritional information in the USDA Nutrient Database for Standard Reference (SR).
Building on this research, we are now designing and developing an automated retail food item classification tool using innovative methods.

Objective
Develop an accurate classification of retail food items by identifying semantically similar concepts in USDA FNDDS records, to support an automated, self-sustaining UPC → FNDDS food code mapping tool.

Background

Data Collection
We obtained grocery transaction data from our retail partner for a sample of 50 households, representing their food shopping activity over a period of 14 months (n=12,332 unique food items)

Data Cleaning
The flat grocery transaction data were cleaned and converted to a normalized relational database.
All USDA FNDDS database tables and the grocery database were imported to a secure MySQL database server using ODBC database connectors.

The Problem
A significant challenge with linking the grocery data was that abbreviated product descriptors were all that was available for the transaction item UPCs.
Like most other retailers, our supermarket data-sharing partner has condensed their descriptions to fewer than 40 characters in order to accommodate the dimensions of an itemized printed sales receipt.

Methods
Web Crawling
• Developed an object-oriented framework in Java that ‘crawls’ the World Wide Web to obtain semantically complete product item descriptors along with their packaging information.
• Used Google’s Search API for Shopping.

Results

Future Enhancements
Fuzzy String matching
• Improve accuracy of matches between our input search parameters from FNDDS and our search results from crawling the Web, using string similarity metrics like the Dice’s coefficient algorithm
• compareStrings(“Cheese”, “Condensed Cheddar Cheese Soup”): 0.35
• compareStrings(“Cheddar cheese soup”, “Condensed Cheddar Cheese Soup”): 0.77

Other methods using NLP:
The iDIET tool will cluster grocery UPC items into semantically meaningful food categories and estimate the semantic relatedness between grocery item names and FNDDS Descriptions for classification.

Conclusion
With its automated data crosswalk in place, iDIET will:
• Facilitate the inclusion of grocery UPCs in dietary studies in a generalizable way.
• Help us explore the feasibility of estimating the overall dietary quality of grocery ‘market baskets’ at the household level, using the USDA’s Healthy Eating Index (HEI)

Acknowledgements
- National Library of Medicine Training Grant No. T15LM007124
- University of Utah, Incentive Seed Grant No: 927

Contact Information
Valliammai Chidambaram
alli.chidambaram@utah.edu