TWENTY QUESTIONS FOR DOCUMENT CLASSIFICATION

George Nagy, Rensselaer Polytechnic Institute, Troy, NY 12180
nagy@ecse.rpi.edu, 1-518-276-6078 (fax: 276-6261)

Sharad Seth, University of Nebraska - Lincoln, Lincoln, NE 68588
seth@cse.unl.edu, 1-402-472-5003 (fax: 472-7767)

Documents – manuscripts, books, magazines, newspapers, sheet music, circuit diagrams, checks, web pages, email attachments, music-CDs, videos, and cuneiform - mirror the culture of the time and serve as the primary source of historical record. Although it seems natural to classify documents according to "format" before examining their content, form and function are often intertwined. The design of a document interpretation system must take both into consideration.

What are the essential parameters of a document interpretation system? What needs to be known before undertaking the design or purchase of such a system? What is the interrelationship of the client, the document, and the desired information? In other words, what is the range of issues of possible interest to our research community? In order to highlight the tacit assumptions implicit in the document analysis literature, we will start with *tabula rasa* and invite the workshop participants to join us in a game of *Twenty Questions*.

The objective of the game is to find out the layout analysis requirements of a partially or fully automated document processing system. We have in mind a realistic application that may stretch the current state of the art. The postulated document analysis system is wanted by a specific organization to extract information from documents of a certain type. The solution will reveal the client's needs and the pertinent layout analysis requirements.

Currently the organization processes 200-500 documents per day through a legacy system where each operator enters data using a workstation networked to a database. The database has ample capacity and a satisfactory query subsystem. Information from a new document must be available on the database within 48 hours of receipt of the document. The current data entry cost averages $1 per document.

You may ask any questions that can be answered by YES or NO. We reserve the option of asking for clarification of the alternatives. Here is a sample of questions that one might ask:

Are the documents *analog* (or *digital)*?
Multimedia (or *mono*)?
Informational or *aesthetic*?
Meant for *eyes* or *ears*?
If *images*, still or video?
  If still, content symbolic or natural?
  If natural, animal, vegetable, or mineral?
  If symbolic, mostly text or mostly graphics?
If *text*, printed or handwritten?
  Is reading order unique?
  Does it contain formulas?
If *structured text*, form or table?
  If form, designed for automated system?
  If table, completely ruled?
Single-page or multi-page?
If *video*, broadcast or low-res?

If *audio*, speech, music, or neither?
  If music, vocal or instrumental?
  If speech, single speaker?
    Structured speech or conversation?
If *graphics*, color or monochrome?
  If monochrome, high-contrast or gray-scale?
  Static or dynamic?
  Pixel array or generative encoding?
  Visually unique?
Searchable?
Editable?
If line drawing, geometric or diagrammatic?
Compressed?
Encrypted?
Does it contain metadata?
Does it contain references?
Active or passive?
Complete or partial conversion required?
Are the required parts identifiable by tags?
  By symbol recognition?
  By format?
Is *ancillary context* information available?
  Is context static or dynamic?
Is there an adequate document model?
Must the *extracted information* be searchable?
  Editable?
  Precisely reproducible?
Is the application unique?
Is a secure system required?
Is the application error tolerant?
Does a long-term need exist?
...

Most current applications, like the one that we have in mind, target only documents of a single type, such as checks, envelopes, insurance claim forms, or *PAMI* articles. But suppose that the organization receives documents of many different types (by mail, parcel post, and the Internet). In this case, it would be necessary to classify each document and route it to the appropriate document processing system. The routing system itself must therefore ask – and answer – some of the questions.

If the questions were fixed ahead of time, it would unduly bind the taxonomy. Also, answers to earlier questions might make some of the later ones meaningless. For example, it would be silly to ask if an audio document is single-page or multi-page. Therefore, as in the *Twenty Questions* game, the system must select questions according to the answers already received. We believe that this game model raises several interesting and serious questions (meta-questions?) about document layout interpretation and its applications.

Q1: What questions should be answerable automatically, at least in principle? Although many of the questions may be well beyond the current state of layout analysis, there is no harm in looking ahead.

Q2: How should the answers be processed?
  Automatically, using
    (a) algorithmic techniques?
    (b) decision trees
    (c) neural-networks
    (d) rule-based expert systems
    (e) autonomous agents
    (f) ...
  Manually, using human experts.

Q3: Can the game be useful even when the document classes are not all known ahead of time? Are there any generic, taxonomy-independent questions? Can we index usefully for future classification needs when entirely unexpected document classes might have emerged?

Q4: What mathematical structure is appropriate to represent the sequential classification game? A tree? A DAG? A lattice?

Q5: What use can we make of current, human-oriented bibliographic tools?

Q6: How can the relationship between layout and content be formalized? Is it possible to classify instruments of knowledge without possessing any knowledge? Can (automated) librarians be illiterate?

Q7. What are the best types of questions? In what sense are they best? When will machine learning and pattern recognition rival human ability to play *Twenty Questions*? Can we hope to stock a classification system with enough questions to play a decent game, or must we instead focus on endowing it with question-making skills? Can such a classifier help further document interpretation?

Q8. How many classes and how many documents might be of interest? How large is the current world population of documents? Is there any fundamental limit to the size of information repositories? Should we advocate stricter document birth control to avoid a Malthusian disaster? Is the information glut only a document glut?

Q9. How would we conduct a document census? Is it enough to concentrate on large-volume archives, such as satellite pictures, newspaper morgues, libraries, hospital records, government offices and insurance companies, or is it important to include "personal" documents like birth/marriage/ death certificates, automobile records, email files, personal libraries? How much would a document census cost? How much would it be worth?

Q10. Can we infer document volumes, by category, from available demographic databases, literacy rates, occupational statistics, library holdings, book/magazine/ music/video outlets, business volume, digital storage device or printer sales? What are the factors that correlate with document production and consumption per capita? What is the average ratio of production to consumption, and how is it changing? How quickly is the world document population shifting from hardcopy to digital? In which domains is the ratio increasing fastest? Slowest?

Q11. What does it all mean for university courses on document image analysis? What about continuing education and research fora? Is there a community of interest? Is there any structure or framework? Are there any facts, observations, theories, which apply to ALL documents, or has the advent of computers rendered the very concept of "document" vacuous?

We beg the indulgence of the participants of the Document Layout Workshop and look forward eagerly to their answers and questions.